Discountinuous Galerkin methods and posteriori error analysis for heterogeneous diffusion problems

Annette Fagerhaug Stephansen

THÈSE

présentée pour l'obtention du titre de

DOCTEUR DE L'ÉCOLE NATIONALE DES PONTS ET CHAUSSÉES

Spécialité : Mathématiques et Informatique

par

Annette Fagerhaug STEPHANSEN

Sujet :

Méthodes de Galerkine discontinues et analyse d'erreur a posteriori pour les problèmes de diffusion hétérogène

pour être soutenue le 17 décembre 2007 devant le jury composé de :

Rapporteurs :	Roland Becker
	Frédéric Pascal
Examinateurs :	Erik Burman
	Philippe Destuynder
	Vivette Girault
	Michel Kern
	Laurent Loth
Directeur de thèse :	Alexandre Ern

Remerciements

Tout d'abord ma gratitude s'adresse à Alexandre Ern pour avoir dirigé ma thèse avec compétence et gentillesse. De ses larges connaissances scientifiques et de sa rigueur, j'ai beaucoup appris.

Ensuite, je souhaiterais remercier l'Andra pour m'avoir permis de réaliser cette recherche et en particulier Alain Dimier pour avoir été mon contact scientifique au sein de cet établissement.

Je tiens aussi à remercier Philippe Destuynder de m'avoir fait l'honneur de présider mon jury de thèse, Roland Becker et Frédéric Pascal d'avoir bien voulu être les rapporteurs de ce travail, Erik Burman et Vivette Girault d'avoir accepté de faire partie de mon jury.

Je voudrais remercier Michel Kern et Laurent Loth pour leurs conseils pendant le travail de thèse et également pour avoir fait partie de mon jury.

Merci à Paolo Zunino de m'avoir accueillie au sein du laboratoire MOX au Politecnico di Milano et pour les discussions fructueuses qui nous ont amenés à la première partie de cette thèse.

Merci à Martin Vohralìk pour les rencontres très enrichissantes. J'ai beaucoup apprécié son enthousiasme et sa rigueur scientifique.

Enfin, merci à toute l'équipe du Cermics de m'avoir accueillie. Les sympathiques discussions après le déjeûner quand tout le monde se réunit autour d'un café vont me manquer! Merci également aux secrétaires Sylvie Berte, Khadija El Louali et Martine Ouhanna pour leur travail d'organisation et leur aide. Un remerciement particulier va à Amélie pour l'organisation des sorties entre thésards et en général pour rendre le milieu du Cermics plus agréable.

Une pensée particulière va à un des mes professeurs du Politecnico di Milano, Luigi Quartapelle, pour m'avoir inspiré sa passion pour les sciences.

Je terminerai cette page en remerciant mes proches qui ont toujours su être là et qui m'ont permis de puiser en moi l'énergie nécessaire pour mener à bien ce travail.

À Vittorio,
À ma famille.

Résumé

Dans cette thèse, nous analysons une méthode de Galerkine discontinue (GD) et deux estimateurs d'erreur *a posteriori* pour l'équation d'advection-diffusion-réaction linéaire et stationnaire avec diffusion hétérogène. La méthode GD considérée, la méthode SWIP, est une variation de la méthode symétrique avec pénalisation intérieure. A la différence de cette dernière, la méthode SWIP utilise des moyennes pondérées dont les poids dépendent de la diffusion. L'analyse *a priori* montre que la convergence est optimale en le pas du maillage et robuste par rapport aux hétérogénéités de la diffusion, ce qui est confirmé par les tests numériques. Les deux estimateurs d'erreur *a posteriori* sont obtenus par une analyse par résidus et contrôlent la (semi-)norme d'énergie de l'erreur. L'analyse d'efficacité locale montre que presque tous les estimateurs sont indépendants des hétérogénéités. L'exception est l'indicateur de non-conformité qui a été évalué en utilisant l'interpolé de Oswald. Le deuxième estimateur d'erreur est plus précis que le premier, mais son coût de calcul est légèrement plus élevé. Cet estimateur est basé sur la construction d'un flux $H(\mathrm{div})$-conforme dans l'espace de Raviart-Thomas-Nédélec en utilisant la conservativité des méthodes GD. Les résultats numériques montrent que les deux estimateurs peuvent être employés pour l'adaptation de maillage.

Abstract

In this thesis we analyse a discontinuous Galerkin (DG) method and two computable a posteriori error estimators for the linear and stationary advection-diffusion-reaction equation with heterogeneous diffusion. The DG method considered, the SWIP method, is a variation of the Symmetric Interior Penalty Galerkin method. The difference is that the SWIP method uses weighted averages with weights that depend on the diffusion. The a priori analysis shows optimal convergence with respect to mesh-size and robustness with respect to heterogeneous diffusion, which is confirmed by numerical tests. Both a posteriori error estimators are of the residual type and control the energy (semi-)norm of the error. Local lower bounds are obtained showing that almost all indicators are independent of heterogeneities. The exception is for the non-conforming part of the error, which has been evaluated using the Oswald interpolator. The second error estimator is sharper in its estimate with respect to the first one, but it is slightly more costly. This estimator is based on the construction of an $H(\mathrm{div})$-conforming Raviart-Thomas-Nédéléc flux using the conservativity of DG methods. Numerical results show that both estimators can be used for mesh-adaptation.

Sommaire

Chapitre 1

Introduction

1.1 Le stockage en formation géologique profonde des déchets radioactifs

La France est un pays pauvre en ressources énergétiques fossiles, ce qui l'a motivée à mettre en œuvre un programme nucléaire important, qui aujourd'hui compte 59 réacteurs. 78 % des kWh éléctriques produits en France sont d'origine nucléaire. Cette production permet de réduire le niveau des émissions de CO_2, mais pose la question des déchets radioactifs. Un des acteurs principaux dans le domaine est l'ANDRA, l'agence nationale pour la gestion des déchets radioactifs. Dans la terminologie du dossier Argile 2005 de l'ANDRA, les déchets HAVL (haute activité, vie longue) comprennent les déchets C (haute activité), B (moyenne activité à vie longue) et CU (combustibles usés non traités). Alors que les déchets CU peuvent être retraités, une solution permanente doit être trouvée pour les déchets B et C, dont le volume total constituait, au 31 décembre 2004, 47 369 m^3.

Un programme de recherche ambitieux a été mis en place avec la loi du 30 décembre 1991, dite 'loi Bataille', qui spécifie trois axes de recherche principaux : la séparation et la transmutation des éléments radioactifs à vie longue, le stockage dans les formations géologiques profondes, et l'étude des procédés de conditionnement et d'entreposage de longue durée en surface. L'ANDRA est responsable de la coordination des recherches sur le deuxième axe. À l'issue des 15 années de recherche menées dans le cadre de la loi de 1991, la loi du 28 juin 2006 renforce le rôle de l'ANDRA en lui confiant les études sur l'entreposage, et établit une feuille de route détaillée pour l'agence. De plus, le stockage en couche géologique profonde devient la solution de référence pour les déchets à haute activité et vie longue.

Un site approprié pour accueillir un stockage doit posséder certaines caractéristiques importantes : le risque sismique à long terme et la présence d'eau doivent être minimaux, la profondeur doit être suffisante (400-700 mètres) pour mettre à l'abri les déchets de diverses perturbations (anthropiques ou naturelles) et la roche doit permettre le creusement pour des installations. La capacité de la roche à limiter la diffusion des déchets radioactifs est cruciale. En outre, des ressources rares exploitables ne doivent pas se trouver à proximité du site. En France, le site actuellement à l'étude est celui de Bure en Meuse/Haute-Marne, où la roche est de type argilite Callovo-Oxfordien.

Un site de stockage souterrain utilise trois différentes barrières pour contenir les déchets radioactifs. La premiere barrière est constituée du conteneur. Les déchets de type B sont compactés ou enrobés dans du bitume ou du béton, avant d'être placés dans un colis en béton ou acier. Les déchets du type C sont incorporés dans un verre particulier et mis dans un conteneur en inox.

Pour la deuxième barrière, dite barrière ouvragée, on a opté pour la bentonite, un type d'argile largement constitué de smectite. La bentonite se gonfle au contact avec l'eau, et, en fonction de sa compacité, peut absorber du liquide correspondant à plusieurs fois son propre poids à sec. C'est une argile très utile pour sceller et rendre un passage imperméable. De plus, la plasticité des argiles leur permet d'absorber des déformations éventuelles.

La roche autour du stockage constitue la troisième barrière, la barrière géologique. L'argilite a une très faible perméabilité et est très homogène. La grande porosité du milieu poreux encourage la fixation des radionucléides à l'intérieur des roches. Un point important est que la zone endommagée créée pendant le creusement des ouvrages soint limitée et n'influence pas les bonnes proprietés du site. Dans le dossier 2005, l'ANDRA a présenté une recherche sur les argiles et l'argilite de Bure en particulier.

Comme les constantes de décroissance radioactive sont de l'ordre de $10^5 - 10^7$ ans pour l'iodine et le plutonium, il est clair que les barrières mises en place n'arriveront pas à contenir toute la radioactivité à l'intérieur du stockage sur de si grandes échelles de temps. L'eau remplira le stockage et les conteneurs seront corrodés. En contact avec l'eau, les radionucléides seront transportés à l'extérieur des alvéoles de stockage vers le milieu naturel. Le risque de contamination de la biosphère est évalué par des analyses de sûreté. Toute l'information disponible est utilisée afin de modéliser les processus, et on se sert du calcul scientifique afin de simuler la migration des radionucléides à un horizon d'un million d'années. Le rôle du groupement de recherches (GDR) MoMaS[1], dans lequel s'inscrit le

[1] http ://www.gdrmomas.org

travail de cette thèse, est de coordonner les recherches sur la modélisation mathématique et la mise au point des schémas numériques pour l'étude et l'analyse des stockages de déchets radioactifs.

De plus amples informations sur le stockage en formation géologique profonde des déchets radioactifs peuvent être trouvées sur le site internet du gouvernement français[2] et sur celui de l'ANDRA[3]. Les dossiers ANDRA sont disponibles au public et peuvent être téléchargés depuis le site internet de l'agence.

1.2 Les milieux poreux et le transport réactif

Un milieu poreux est constitué d'une structure solide et d'un réseau de pores à travers lesquels un fluide s'écoule. L'ensemble formé de la structure, des pores et du fluide peut être considéré comme un continuum. La modélisation et l'étude des milieux poreux sont très importantes, car elles permettent de comprendre le comportement de nombreux matériaux. Même si le concept de milieu poreux a été originellement développé pour étudier la mécanique des sols, et donc pour caractériser le sable ou les roches, on a ensuite utilisé le concept pour former une théorie plus générale, la poromécanique. La caractérisation des milieux poreux est donc utilisée aussi pour des matériaux naturels comme le bois et les tissus biologiques et des matériaux industriels comme la céramique et la mousse.

On caractérise un milieu poreux par sa porosité, sa perméabilité et par les propriétés de sa structure solide et du fluide qui la traverse. La porosité est définie par le rapport entre le volume des pores et le volume total, et est donc par définition plus petite que 1. La perméabilité dépend exclusivement de la géométrie du milieu et mesure la possibilité du fluide de le traverser. Elle dépend non seulement de la porosité du milieu, mais aussi de la connexité des pores, et est exprimée par le biais d'un tenseur symétrique défini positif. La perméabilité peut être mesurée expérimentalement, et pour les milieux anisotropes les mesures doivent être faites en considérant les trois directions spatiales. Dans le cas général d'écoulements dans les roches, une caractérisation complète doit en outre considérer les fractures présentes.

L'écoulement dans un milieux poreux est souvent modélisé par l'équation de Darcy

$$q = -K\nabla h, \tag{1.1}$$

[2] http ://www.industrie.gouv.fr/energie/sommaire.htm
[3] http ://www.andra.fr

où q est la vitesse de filtration, K est la conductivité hydraulique et h est la charge hydraulique. Le signe négatif est dû au fait que la filtration a lieu dans la direction des pressions décroissantes. La conductivité hydraulique dépend de la perméabilité et de la viscosité dynamique du fluide. La loi de Darcy peut être déduite des équations de Navier-Stokes sous certaines hypothèses. En fait, l'équation (1.1) est valable seulement pour des écoulements lents et visqueux, ce qui est le cas pour le transport de solutés dans un sol. Par contre, si le transport est polyphasique, l'équation doit être modifiée. Si le flux est monophasique et stationnaire et sous l'hypothèse de densité constante, la conservation de la masse en absence de sources ou puits implique que

$$\nabla \cdot q = 0. \tag{1.2}$$

Les équations (1.1) et (1.2) completées des conditions aux limites décrivent, à l'échelle macroscopique, l'écoulement stationnaire dans un sol saturé.

Pour décrire le transport d'une espèce radioactive de concentration C_i dans un sol, on utilise en première approximation l'équation d'advection-diffusion-réaction suivante

$$\phi R_i \left(\frac{\partial C_i}{\partial t} + \lambda_i C_i \right) - \nabla \cdot (D_i \nabla C_i) + q \cdot \nabla C_i = f_i,$$

où ϕ est la porosité effective, R_i est le facteur de retard, λ_i la constante de décroissance radioactive, q la vitesse de Darcy, D_i le tenseur de diffusion/dispersion et f_i un terme de source ou puits éventuel. L'indice i indique que la valeur peut être différente pour chaque espèce. La porosité effective fait référence aux pores qui sont ouverts pour le transport du fluide.

Le facteur de retard tient compte du fait que le soluté n'est pas transporté avec la même vitesse que le fluide. C'est une manière très simple de modéliser l'effet du processus chimique d'adsorption, qui par contre est très compliqué. L'adsorption fait intervenir la surface du milieu poreux (à la différence de l'absorption où les molécules sont incorporées dans la structure même). L'adsorption est souvent modélisée comme un processus instantané en utilisant des isothermes. Celles-ci fournissent la masse adsorbée S sur la surface en fonction de la concentration du soluté C. Les deux isothermes les plus utilisées sont celles de Langmuir et de Freundlich. L'isotherme de Freundlich est donnée par

$$S = K[C]^n,$$

alors que l'isotherme de Langmuir est donnée par

$$S = \frac{KMC}{1 + KC},$$

où M est le maximum de soluté adsorbé. Le facteur de retard est calculé en utilisant la relation

$$R = 1 + \rho \frac{1-\phi}{\phi} K_d,$$

où ρ est la densité de la phase solide et $K_d = S/C$ est le coefficient de distribution. Notons que le facteur de retard peut donner lieu à d'importantes non linéarités.

La modélisation des phénomènes de diffusion par le terme $\nabla \cdot (D_i \nabla C_i)$ est connue sous le nom de 'loi de Fick', et reste acceptable quand un intervalle de temps relativement grand est considéré. Le tenseur de diffusion/dispersion D_i tient compte de différents phénomènes, notamment la diffusion moléculaire et la diffusion mécanique. Le tenseur peut être décrit en utilisant la relation suivante

$$D_i = d_{ei} I + \alpha_{li} F(q) + \alpha_{ti}(I - F(q)).$$

La diffusion moléculaire est due au mouvement brownien, et rend compte du mouvement du soluté au niveau moléculaire. Elle est modélisée par le terme $d_{ei}I$, où I est la matrice identité. La valeur de d_{ei} est en général très petite. La dispersion mécanique est due à la variation locale de la vitesse par rapport à la vitesse macroscopique (de Darcy), et est modélisée par deux coefficients α_{li} et α_{ti} et un tenseur F qui dépend de la vitesse q. La diffusion mécanique rend souvent négligeable la diffusion moléculaire, sauf quand la vitesse d'écoulement devient très petite comme dans l'argilite.

Le système d'équations d'advection-diffusion-réaction pour les différentes concentrations des espèces chimiques doit être accompagné d'un système d'équations qui décrit l'interaction entre celles-ci. Les processus à considérer sont les phénomènes en phase aqueuse, les échanges liquide-gaz et les échanges liquide-solide. Il faut aussi considérer la cinétique des réactions. Dans ce travail de thèse, nous nous sommes limités à considérer la résolution de l'équation d'advection-diffusion-réaction stationnaire et linéaire.

Pour plus d'informations sur le transport de contaminants, on pourra consulter le livre d'Anderson [7].

1.3 Les méthodes de Galerkine discontinues

Contrairement aux méthodes usuelles d'éléments finis, les méthodes de Galerkine discontinues (GD) n'imposent pas de contrainte de continuité sur les fonctions de base, ce qui conduit à une solution approchée qui n'est pas H^1-conforme.

La définition de l'espace d'approximation GD sur un maillage $\mathcal{T}_h$ du domaine Ω, supposé polygonal ou polyédrique pour simplifier, est

$$V_h = \{v_h \in L^2(\Omega); \forall T \in \mathcal{T}_h, v_h|_T \in \mathbb{P}_p\},$$

où $\mathbb{P}_p$ est l'ensemble de polynômes de degré global inférieur ou égal à p. Les méthodes GD donnent lieu à la formulation suivante du problème approché : Chercher $u_h \in V_h$ tel que

$$a_h(u_h, v_h) = (f, v_h)_{0,\Omega}, \qquad \forall v_h \in V_h,$$

où f est la donnée du problème, $(\cdot,\cdot)_{0,\Omega}$ le produit scalaire L^2 sur le domaine Ω et a_h la forme bilinéaire de la méthode GD. La forme bilinéaire a_h contient non seulement les termes rencontrés dans les méthodes d'éléments finis usuelles, mais également d'autres termes. Ces termes diffèrent de méthode en méthode, mais il y a normalement un terme qui rend la méthode consistante et un autre dit de pénalisation, afin d'imposer de manière faible la continuité de la solution approchée et les conditions aux limites. Le terme de consistance peut également être accompagné d'un terme qui rend la matrice de rigidité symétrique.

Les méthodes GD peuvent aussi être formulées en termes de flux numériques définis sur les interfaces d'un élément. Les flux sont dits conservatifs si leur valeur est unique au signe près sur une face partagée par deux éléments. Dans ce cas, le flux qui sort d'un élément est égal au flux qui entre dans l'élément voisin. Cette propriété est commune avec les méthodes de volumes finis.

Comme avec les méthodes d'éléments finis usuelles, la matrice de rigidité obtenue avec une méthode GD est creuse, et il est possible de travailler sur des géométries complexes en utilisant des maillages non structurés. La non-conformité de la méthode donne en outre la possibilité de décomposer le domaine en sous-domaines, et de mailler ceux-ci séparément sans contraintes de compatibilité entre les maillages des sous-domaines. Les nœuds pendants ('hanging nodes' en anglais) sont donc autorisés, et il est possible de raffiner le maillage ou d'augmenter localement le degré des polynômes utilisés sans grande difficulté. Enfin, travailler avec des fonctions de base discontinues peut sembler naturel si des couches limites sont présentes dans la solution exacte.

L'inconvénient principal des méthodes GD est leur nombre de degrés de liberté élevé par rapport à celui des méthodes d'éléments finis usuelles. Il est donc important d'exploiter les avantages de la méthode au regard de la flexibilité des maillages afin de minimiser les coûts du calcul.

Les méthodes GD ont été introduites par Reed et Hill [83] dans les années 70 afin de résoudre un problème hyperbolique lié au transport de neutrons. Lesaint et Raviart [69]

ont effectué l'analyse mathématique de la méthode en 1974, mais la période la plus propice au développement des méthodes GD pour les problèmes hyperboliques a été vers la fin des années 80 et les années 90. Avec la première conférence internationale sur les méthodes GD qui a eu lieu en 1999, le développement et les problématiques liés à ces méthodes ont été tracés, voir par exemple l'ouvrage de Cockburn, Karniadakis et Shu [31]. Grâce au fait que les méthodes GD pour les systèmes hyperboliques non-linéaires ont été formulées en termes de flux numériques, les similarités avec les méthodes de volumes finis ont pu être exploitées. Les façons de stabiliser le schéma et de capturer les discontinuités éventuelles dans la solution exacte sont souvent très similaires. Les méthodes implicites SCDG (Shock Capturing Discontinuous Galerkin) ajoutent un terme de viscosité artificielle qui dépend du résidu de l'équation et de la taille locale h du maillage ; cf. Jiang et Shu [60]. La démonstration que la convergence du schéma est d'ordre élevé est facilitée par la dépendance de ce terme en h, mais à proximité des discontinuités la viscosité artificielle atténue trop les extrema de la solution. Les méthodes RKDG (Runge Kutta Discontinuous Galerkin) ont été introduites par Cockburn et Shu [32]. Dans ces méthodes, la viscosité numérique dépend de la régularité locale de la solution. Elles peuvent être réécrites avec un limiteur de pente, utilisé pour avancer la solution en temps après avoir résolu l'équation sans viscosité artificielle dans les pas de temps intermédiaires.

Les méthodes GD sont également en mesure d'approcher les systèmes de lois de conservation comportant des dérivées du deuxième ordre. En étudiant les équations de Navier-Stokes, Bassi et Rebay [16] ont décidé de traiter également comme inconnue la dérivée première de la solution. Le système a ensuite été résolu avec une méthode RKDG. Cockburn et Shu [33] ont généralisé cette procédure en proposant la méthode LDG (Local Discontinous Galerkin). Dans le schéma LDG, toutes les équations sont récrites comme un système d'équations du premier ordre. C'est donc la formulation mixte du problème de départ qui est considérée.

Avec le développement des méthodes RKDG et LDG, les méthodes GD sont devenues de première importance dans la résolution des systèmes de lois de conservation avec termes du premier et du deuxième ordre. Outre les équations de Navier-Stokes, nous pouvons mentionner à titre d'exemple la résolution des équations de Saint Venant par Ern, Piperno et Djadel [47], par Tassi, Bokhove et Vionnet [95] et par Aizinger et Dawson [6].

Le développement des méthodes GD pour les problèmes elliptiques s'est fait de manière relativement indépendante, en s'inspirant de la formulation faible des conditions aux limites proposées dans les travaux de Nitsche [76, 77]. Les premiers travaux ont été effectués dans les années 70-80, et sont ceux de Babuška [11], de Babuška et Zlámal [14], de Douglas et

Dupont [40], de Baker [15], de Wheeler [104] et de Arnold [10]. Dans ces travaux, les flux numériques ne sont pas exprimés de manière explicite, et ce sont plutôt les modifications possibles des différents termes de la forme bilinéaire qui sont considérées.

La méthode GEM (Global Element Method) de Delves et Hall [36] est une méthode sans termes de pénalisation pour laquelle la matrice de rigidité est symétrique. Comme la matrice n'est pas nécessairement semi-définie positive, la méthode peut être inconditionnellement instable en l'appliquant à une équation instationnaire. Il n'a pas non plus été démontré que le schéma soit bien posé. La méthode hp DG (hp Discontinuous Galerkin) de Oden, Babuška et Baumann [78] se différencie de la méthode précédente par un signe, ce qui rend les termes de consistance antisymétriques. Pour obtenir un schéma stable, il faut par contre utiliser des polynômes d'un degré minimum de deux.

Parmi les schémas de type IP (Interior Penalty - pénalisation interne) on trouve la méthode constituant le point de départ de cette thèse : la méthode SIPG (Symmetric Interior Penalty Galerkin) établie suite aux travaux de Baker [15], de Douglas et Dupont [40], de Wheeler [104] et de Arnold [10]. La forme bilinéaire est symétrique, et les sauts de la solution approchée ainsi que les conditions aux limites de Dirichlet sont pénalisés. Le paramètre de stabilisation doit être supérieur à un certain seuil minimal qui doit être déterminé par l'utilisateur. Une variante de la méthode SIPG qui s'affranchit de ce paramètre indéterminé, est la méthode proposée par Bassi, Rebay, Mariotti, Pedinotti et Savini [17], où le terme de pénalisation utilise des opérateurs de relèvement. La méthode NIPG de Rivière, Wheeler et Girault [89] est très similaire, modulo un des termes qui est changé de signe. Dans ce cas, on obtient une méthode qui assure la positivité de la forme bilinéaire sans le terme de stabilisation. Par ailleurs, un désavantage de la méthode NIPG est que, contrairement à la méthode SIPG, on ne sait pas montrer la convergence optimale de l'erreur en norme L^2 sous hypothèse de régularité elliptique, même si cette convergence est observée dans les essais numériques. Une variante de la méthode NIPG a été présentée par Romke, Oden et Prudhomme [92], où sont pénalisés les sauts des flux diffusifs et non les sauts de la solution approchée.

Avec l'introduction des méthodes LDG, les similarités entre les méthodes DG pour les équations hyperboliques et celles pour les équations elliptiques sont devenues plus évidentes. Dans les dix dernières années, les analyses unifiées des méthodes GD ont vu le jour. Mentionnons l'article de Arnold, Brezzi, Cockburn et Marini [8] qui a pour but d'unifier l'analyse des méthodes GD appliquées aux équations elliptiques. Les articles de Ern et Guermond [43–45] et de Di Pietro, Ern et Guermond [38], par ailleurs, examinent les méthodes GD pour les systèmes de Friedrichs, qui comprennent à la fois les équations

hyperboliques et elliptiques.

Pour le cas qui nous intéresse dans cette thèse, les équations d'advection-diffusion-réaction, l'analyse des méthodes GD a été présentée de manière approfondie dans l'article de Houston, Schwab et Süli [59], qui couvre notamment le cas d'un tenseur de diffusion anisotrope et hétérogène. Par contre, le cas particulier d'une très petite diffusion dans une partie du domaine pose encore des difficultés. En effet, dans le cas où le champ d'advection est orienté dans la direction de diffusion (isotrope) croissante, une couche limite se forme à l'interface où la diffusion est discontinue. Dans le cas limite de diffusion évanescente d'un côté de l'interface, la solution exacte est discontinue si le champ advectif pointe de la partie hyperbolique vers la partie elliptique. Ce cas a été analysé en une dimension d'espace par Gastaldi et Quarteroni [52], et plus récemment dans l'article de Croisille, Ern, Lelièvre et Proft [34]. Dans le cas d'une diffusion anisotrope en dimension ≥ 2, il faut tenir compte de la diffusion et de l'advection dans la direction normale à l'interface, cf. Di Pietro, Ern et Guermond [38].

Lorsque certaines des valeurs propres du tenseur de diffusion deviennent très petites, même si le tenseur reste défini positif, les méthodes GD usuelles ont des difficultés si la couche limite n'est pas suffisamment résolue par le maillage. En effet, au lieu d'imposer la continuité d'une manière faible, la solution (continue) serait mieux approchée par une fonction discontinue sur l'interface où se trouve la couche limite. Une possibilité serait d'autoriser cette discontinuité dans la programmation de la méthode en éliminant manuellement les terms de pénalisation sur l'interface en question, ainsi que proposé par Houston, Schwab et Süli [59] et plus récemment dans l'article de Ern et Proft [48].

Notre proposition, qui sera décrite en détail par la suite, consiste en revanche à modifier les méthodes GD de façon plus générale. Dans les termes de consistance, nous proposons de considérer des moyennes pondérées au lieu des moyennes arithmétiques, avec des poids dépendant de la diffusion. Le terme de pénalisation, par ailleurs, dépend de la moyenne harmonique de la diffusion dirigée normalement à l'interface. La seule hypothèse (raisonnable) qui est requise, est que les discontinuités dans la diffusion coïncident avec certaines des interfaces du maillage, ce qui est raisonnable dans le contexte de la modélisation hydrogéologique.

L'utilisation des moyennes pondérées provient de la méthode d'éléments finis dite de 'mortier', dont l'idée remonte aux travaux de Nitsche [76, 77]. Cette méthode impose la continuité des flux entre régions différentes de manière faible. Divers auteurs ont noté la possibilité d'utiliser une moyenne avec des poids dans les méthodes GD. Mentionnons les travaux de Stenberg [94], de Heinrich et Pönitz [56], de Heinrich et Nicaise [54] et

de Heinrich et Pietsch [55]. Dans ces travaux, différentes techniques de type 'mortier' ont été proposées afin d'utiliser des éléments finis conformes sur un maillage qui n'est pas nécessairement conforme. Les moyennes pondérées sont introduites simplement pour généraliser les moyennes arithmétiques. Par contre, les poids ne sont pas choisis en fonction des coefficients du problème et notamment du coefficient de diffusion. Un tel choix a été récemment exploré dans l'article de Burman et Zunino [25] pour des problèmes d'advection-diffusion-réaction isotrope approchés par une technique de type 'mortier'. Si on applique cette méthode élément par élément, on obtient une méthode GD. Burman et Zunino ont montré qu'un choix spécifique des poids améliore la stabilité du schéma quand la diffusivité prend localement des valeurs très petites. L'extension à des équations d'advection-diffusion-réaction avec diffusion localement évanescente a été analysée récemment par Di Pietro, Ern et Guermond [38].

1.4 Analyse d'erreur *a posteriori*

L'analyse d'erreur *a priori* vise à démontrer la bonne convergence du schéma numérique ; l'analyse d'erreur *a posteriori* a pour but de quantifier l'erreur d'approximation. Celle-ci doit être mesurée dans une norme qui est significative pour le problème en question : pour l'évaluer, nous avons à disposition la solution calculée, les données du problème et les données du maillage. Dans cette thèse, nous nous restreindrons à des estimateurs d'erreur *a posteriori* dans la norme de stabilité naturelle du problème continu, ou (semi-)norme d'énergie, que nous noterons $\|\cdot\|_B$.

Un estimateur d'erreur doit fournir une borne supérieure de la véritable erreur. Nous nommerons u la solution exacte et u_h la solution GD. Ainsi, $e(h, f, u_h)$ est un estimateur d'erreur si

$$\|u - u_h\|_B \leq e(h, f, u_h),$$

où h est la taille du maillage et f le terme source. Définissons l'indice d'efficacité I_e par

$$I_e = \frac{e(h, f, u_h)}{\|u - u_h\|_B}.$$

Afin de pouvoir décider si le calcul effectué a été suffisamment précis, l'estimateur $e(h, f, u_h)$, accessible par le calcul, est évalué. Un indice d'efficacité proche de 1 implique que l'erreur n'est pas inutilement surévaluée. De plus, il est souhaitable que l'indice d'efficacité soit indépendant des données du problème (comme le tenseur de diffusion ou le champ d'advection). Dans ce cas, l'estimateur est dit robuste.

L'adaptation de maillage est un bon moyen pour obtenir une solution plus précise avec un surcoût de calcul modéré, un facteur à ne pas négliger si les calculs sont d'une certaine importance. Un ingénieur avec beaucoup d'expérience peut utiliser son intuition pour identifier les parties du domaine où la méthode aurait besoin d'un maillage plus fin. Une autre procédure relativement courante consiste à raffiner le maillage dans les parties où la solution calculée présente un fort gradient. Par opposition, l'analyse d'erreur *a posteriori* vise à automatiser la procédure d'adaptation du maillage en basant cette procédure sur des fondements mathématiques solides.

Afin de pouvoir utiliser l'estimation d'erreur *a posteriori* pour adapter le maillage, il faut que l'estimateur soit localisable, c'est-à-dire que l'estimateur puisse s'écrire comme une somme sur les éléments du maillage $\mathcal{T}_h$ du domaine Ω sous la forme

$$e(h, f, u_h) = \left(\sum_{T \in \mathcal{T}_h} e_T^2(h, f, u_h) \right)^{\frac{1}{2}}.$$

Les quantités $e_T(h, f, u_h)$ sont appelées indicateurs d'erreur. Dans la partie du domaine où les indicateurs sont les plus grands, le maillage est raffiné. Il est aussi possible de déraffiner le maillage dans les parties du domaine où les indicateurs sont les plus petits. Dans l'algorithme d'adaptation, il est possible, par exemple, de raffiner les éléments où l'indicateur d'erreur dépasse l'indicateur le plus grand multiplié par une constante $c < 1$, et de déraffiner suivant un critère similaire. Une autre possibilité est de décider *a priori* le pourcentage d'éléments à raffiner pour mieux contrôler le coût de calcul d'un maillage à l'autre. Une troisième possibilité est d'utiliser le marquage proposé par Dörfler [39] qui consiste à trouver un sous-ensemble minimal de mailles dont la contribution des indicateurs représente une fraction minimale de l'estimateur total. Cette stratégie de marquage permet de garantir la réduction de l'erreur sous certaines hypothèses, voir également les travaux de Morin, Nochetto et Siebert [71–73].

Pour que les indicateurs d'erreur soient utiles, il faut qu'il ne surestiment pas trop l'erreur localement. Dans le cas idéal, l'indicateur et la véritable erreur sont localement équivalents, c'est à dire, sur chaque élément $T \in \mathcal{T}_h$,

$$c_1 e_T(h, f, u_h) \leq \|u - u_h\|_{B,T} \leq c_2 e_T(h, f, u_h). \tag{1.3}$$

Ici c_1 et c_2 sont deux constantes, et la norme indiquée par B, T est telle que

$$\|u - u_h\|_B = \left(\sum_{T \in \mathcal{T}_h} \|u - u_h\|_{B,T}^2 \right)^{\frac{1}{2}}.$$

En général, le mieux que l'on puisse obtenir est une majoration de l'indicateur local par l'erreur locale en considérant les éléments les plus proches, c'est-à-dire :

$$c_1 e_T(h, f, u_h) \leq \sum_{T \in \Delta_T} \|u - u_h\|_{B,T},$$

où Δ_T indique un ensemble d'éléments autour de T, par exemple les éléments partageant au moins une face avec T.

Les indicateurs d'erreur ont été introduits dans les années 70 par Babuška et Rheinboldt [12, 13]. Plusieurs techniques ont été développées par la suite. Nous nous concentrerons ci-après sur les estimateurs par résidu, car ils constituent un des principaux sujets de cette thèse.

Considérons par exemple l'équation de Poisson avec conditions aux limites de Dirichlet homogènes sur un domaine Ω en $\mathbb{R}^d$ de frontière $\partial\Omega$:

$$\begin{cases} -\Delta u = f & \text{sur } \Omega, \\ u = 0 & \text{sur } \partial\Omega. \end{cases} \tag{1.4}$$

Indiquant par u_h la solution approchée obtenue avec une méthode GD, le résidu est égal à

$$R(u_h) = f + \Delta_h u_h, \tag{1.5}$$

où Δ_h indique le laplacien local, c'est-à-dire l'opérateur qui sur chaque élément coïncide avec l'opérateur Δ appliqué à la restriction sur cet élément. Dans l'estimateur d'erreur par résidu, le terme $R(u_h)$ est accompagné d'autres termes, typiquement un terme qui mesure la non-conformité de u_h et d'autres termes qui mesurent les sauts des flux diffusifs et les sauts de u_h sur les interfaces.

Pour l'équation de Poisson avec conditions aux limites de Dirichlet, la semi-norme d'énergie est la norme L^2 du gradient brisé, c'est à dire le gradient défini maille par maille. Pour les méthodes GD, les premières estimations d'erreur par résidu dans la semi-norme d'énergie ont été obtenues par Becker, Hansbo et Larson [20] et par Karakashian et Pascal [62]. Ainsworth [3, 4] a rendu explicite la dépendance des constantes vis à vis de la diffusion, tandis que Houston, Schötzau et Wihler [58] ont effectué une analyse *hp*. En ce qui concerne les estimateurs d'erreur en norme L^2, on peut mentionner le travail de Becker, Hansbo et Stenberg [21], celui de Rivière et Wheeler [87] et celui de Castillo [29].

Dans les travaux de Becker, Hansbo et Larson [20], Ainsworth [3] et Castillo [26], le gradient de l'erreur $e = u - u_h$ est sujet à une décomposition de Helmholtz ; une technique introduite à l'origine dans les articles de Dari, Duran, Padra, et Vampa [35] et Carstensen,

Bartels et Jansche [27]. La technique consiste à considérer que ∇e est composé de deux parties :

$$\nabla e = \nabla\phi + \nabla \times \varphi,$$

où $\phi \in H_0^1(\Omega)$ est un potentiel scalaire tel que

$$(\nabla\phi, \nabla v)_{0,\Omega} = \sum_{T\in\mathcal{T}_h} (\nabla e, \nabla v)_{0,T}, \qquad \forall v \in H^1(\Omega),$$

et $\varphi \in \mathcal{H} = \{H^1(\Omega), (\varphi, 1)_{0,\Omega} = 0\}$ est un potentiel vecteur tel que

$$(\nabla \times \varphi, \nabla \times w)_{0,\Omega} = \sum_{T\in\mathcal{T}_h} (\nabla e, \nabla \times w)_{0,T}, \qquad \forall w \in \mathcal{H}.$$

La décomposition conduit à l'identité suivante pour la norme d'énergie

$$\sum_{T\in\mathcal{T}_h} \|\nabla e\|_{0,T}^2 = \sum_{T\in\mathcal{T}_h} \|\nabla\phi\|_{0,T}^2 + \sum_{T\in\mathcal{T}_h} \|\nabla \times \varphi\|_{0,T}^2.$$

L'analyse s'effectue en intégrant par parties et en notant que $\nabla\phi = \nabla(\phi - \pi\phi)$ où $\pi\phi$ est la projection L^2-orthogonale sur l'espace des fonctions constantes par morceaux. Cette façon de procéder a aussi été appliquée aux équations de Maxwell par Houston, Perugia et Schötzau [57]. Par contre, elle n'a pas été utilisée dans l'analyse de Houston, Schötzau et Wihler [58], ni dans celle de Karakashian et Pascal [62]. Nous ne n'utilisons pas non plus dans cette thèse. Notons que la technique de décomposition de Helmholtz ne s'étend pas facilement si la norme d'énergie dans laquelle on souhaite contrôler l'erreur contient des termes d'ordre zéro, comme c'est le cas pour les équations d'advection-diffusion-réaction.

Dans tous les cas, un terme de non-conformité est présent dans les estimations d'erreur *a posteriori* pour les méthodes GD. D'une façon générale, ce terme peut être formulé en introduisant une fonction continue arbitraire qui doit respecter les conditions aux limites de Dirichlet. Afin de pouvoir calculer cette erreur de non-conformité, il faut choisir une fonction spécifique, et l'interpolé de Oswald est un choix courant. Sur chaque nœud du maillage qui n'est pas situé sur la frontière, l'interpolé de Oswald prend la valeur moyenne de la solution calculée u_h. Si le nœud est à l'intérieur de l'élément, u_h et son interpolé de Oswald prennent la même valeur. Si le nœud se trouve sur l'interface entre deux éléments, la valeur de l'interpolé de Oswald est la moyenne arithmétique des deux valeurs de u_h, etc. Les valeurs au sein de chaque élément sont ensuite interpolées avec des polynômes de lagrange.

Une des difficultés avec les estimateurs ci-dessus est que pour le problème de Poisson, si la solution u_h est linéaire par morceaux, le résidu est identique au terme source f du problème. Cette estimation est trop grossière pour les méthodes GD, qui disposent de tous les degrés de liberté polynomiaux dans chaque maille, si bien qu'on devrait obtenir des résidus du type $\|f - \pi_p f\|_{0,T}$ où $\pi_p f$ indique la projection L^2 orthogonale du terme source sur l'espace vectoriel des polynômes de degré p. Les estimateurs d'erreur obtenus récemment par Vohralík pour les méthodes mixtes et de volumes finis [101–103] considèrent notamment des résidus de ce type. Pour les méthodes de volumes finis, on a $p = 0$, alors que pour les méthodes mixtes, p est le degré polynomial de l'inconnue scalaire. Ce résultat a été obtenu en utilisant le fait que les méthodes mixtes et de volumes finis sont localement conservatives. Notre contribution a été d'étendre ce resultat aux méthodes GD. Notons que si f est suffisamment régulière (i.e. $f \in H^1(T)$ pour tout $T \in \mathcal{T}_h$) la convergence du résidu est d'un ordre plus élevé par rapport au résidu standard.

Dans les estimateurs par résidu, le terme qui mesure le défaut de conservativité des flux diffusifs basés sur le gradient local de la solution calculée, résulte du fait que le gradient n'est pas dans l'espace $H(\mathrm{div}, \Omega) = \{v \in L^2(\Omega); \nabla \cdot v \in L^2(\Omega)\}$. Ce terme n'est pas strictement local, car pour le calculer, il faut considérer les éléments qui partagent une interface avec un élément donné. Il serait donc souhaitable de pouvoir substituer à ce terme un autre qui serait calculé en considérant seulement l'élément en question. Même si les flux diffusifs ne sont pas continus, les flux numériques des méthodes GD le sont. Il est donc envisageable d'utiliser les flux numériques afin de construire un champ vectoriel dans $H(\mathrm{div}, \Omega)$, et ensuite de baser l'estimation d'erreur sur ce champ vectoriel. Pour que cette technique soit intéressante, la construction du champ vectoriel devra être locale, ce qui limite en effet le coût du calcul. Notre contribution porte également sur ce point.

L'idée d'utiliser une construction de flux $H(\mathrm{div}, \Omega)$-conformes dans les estimations d'erreur remonte aux années 40 avec le travail de Prager et Synge [81]. Pour l'application aux méthodes d'éléments finis conformes, signalons les travaux de Ladevèze [65], de Ladevèze et Leguillon [66], de Destuynder et Métivet [37], de Repin [84–86] et de Neittaanmäki et Repin [75]. L'application aux méthodes GD pour les problèmes de diffusion pure est par contre très récente, et a été explorée par Ainsworth [5], par Kim [63, 64], par Lazarov, Repin et Tomar [67] et par Cochez-Dhondt et Nicaise [30]. L'application aux problèmes d'advection-diffusion-réaction est nouvelle, et sera explorée ci-après. Une observation importante est que dans tous ces estimateurs, il n'y a pas de constantes indéterminées.

1.5 Objectifs de la thèse

Le principal objectif de cette thèse est d'améliorer la résolution de l'équation d'advection-diffusion-réaction linéaire et stationnaire, dans le cas où le tenseur de diffusion présente de fortes hétérogénéités. L'application visée est la modélisation de la dispersion de composants radioactifs autour d'un ouvrage de stockage souterrain dans le milieu naturel. En particulier, cette dispersion est modélisée par une équation de transport réactif simplifiée en milieu poreux. Les différentes couches rocheuses sont caractérisées par des diffusions très hétérogènes, ce qui pose une difficulté pour les méthodes numériques. Vu les propriétés favorables des méthodes GD, nous nous sommes concentrés sur cette famille de schémas numériques.

Le premier objectif est de construire une méthode d'approximation robuste et précise. Nous supposons que la diffusion est constante à l'intérieur de chaque élément si bien que les discontinuités dans le tenseur de diffusion coïncident avec des frontières de certains éléments du maillage. En modifiant la méthode SIPG standard, nous obtenons un schéma qui impose moins de continuité sur les interfaces où les hétérogénéités donnent lieu à des forts gradients. En permettant une discontinuité plus forte dans la solution calculée par rapport à celle imposée par la méthode SIPG usuelle, nous arrivons à diminuer, voire en certain cas à éliminer, les oscillations qui sont souvent présentes à proximité d'une couche limite. Le résultat obtenu est dû à un choix des poids utilisés dans le terme de consistance et dans le terme qui rend la méthode symétrique. Dans la méthode SIPG, les poids sont tout simplement égaux à $\frac{1}{2}$. Le choix du paramètre de pénalisation est aussi très important, et nous utilisons la moyenne harmonique de la diffusion dans la direction normale à l'interface, alors que les méthodes SIPG proposées dans la littérature utilisent souvent la moyenne arithmétique. Dans l'analyse nous verrons que la possibilité d'utiliser la moyenne harmonique est une conséquence du choix des poids dans le terme de consistance. Nous avons nommé cette nouvelle méthode SWIP pour Symmetric Weighted Interior Penalty.

La qualité du maillage est cruciale pour garantir des résultats numériques satisfaisants, en particulier si la solution présente des couches limites. Pour l'analyse de sûreté, il est aussi très important de pouvoir se fier aux résultats numériques obtenus. Le but sera donc d'effectuer l'analyse d'erreur *a posteriori* pour obtenir un estimateur d'erreur qui puisse également être utilisé pour l'adaptation du maillage. Nous analysons dans cette thèse deux estimateurs d'erreur basés sur l'analyse par résidus. Ces estimateurs sont facilement calculables, et fournissent des informations sur la semi-norme d'énergie (à savoir la norme

L^2 du gradient brisé) de l'erreur dans tout le domaine. On observera que le terme de sauts aux interfaces et les valeurs au bord de la solution approchée, qui sont souvent inclus dans les estimateurs d'erreur *a posteriori* pour les méthodes GD, ne sont pas inclus dans la semi-norme d'énergie car ces termes dépendent du choix des paramètres numériques de la méthode. L'analyse d'erreur se base sur la possibilité d'identifier la non-conformité de la solution, c'est-à-dire l'erreur commise en utilisant des fonctions discontinues pour approcher une solution continue. Le premier estimateur que nous avons obtenu est similaire dans sa forme à celui de Karakashian et Pascal [62]. Nous avons par ailleurs apporté beaucoup de soin à l'évaluation de toutes les constantes et à l'amélioration de l'efficacité de l'estimateur afin de le rendre le plus indépendant possible des hétérogénéités du tenseur de diffusion. Le deuxième estimateur utilise un champ vectoriel auxiliaire construit par le biais de problèmes locaux. La procédure est inspirée des travaux de Vohralík pour les méthodes mixtes [102] et de volumes finis [101], et est basée sur la propriété de conservativité des méthodes GD. Le coût de calcul du deuxième estimateur est légèrement plus grand, mais en revanche son indice d'efficacité est meilleur.

1.6 Plan de la thèse

Ce mémoire est composé de 5 chapitres.

Dans le chapitre 2 nous présentons l'équation d'advection-diffusion-réaction considérée et la méthode SWIP, notamment le choix des poids et du paramètre de stabilisation. Nous montrons que la méthode proposée est coercive par rapport à la norme d'énergie du problème discret et que la convergence en norme d'énergie et en norme L^2 (sous hypothèse de régularité elliptique) est d'ordre optimal. De plus, cette convergence est indépendante des hétérogénéités et des anisotropies du tenseur de diffusion. Afin de compléter l'analyse de convergence de la méthode dans le cas de diffusion évanescente, nous effectuons aussi l'analyse d'erreur par rapport à la dérivée advective. Ce résultat est lui aussi indépendant des hétérogénéités du tenseur de diffusion, mais l'anisotropie du tenseur peut affecter l'erreur dans certains cas. La robustesse de l'estimation d'erreur est obtenue si les nombres de Péclet évalués par rapport à la plus grande valeur propre du tenseur de diffusion local sont suffisamment grands. Enfin nous présentons des tests numériques pour illustrer l'analyse d'erreur. Les mêmes tests numériques ont été réalisés avec la méthode SIPG afin de comparer les résultats avec ceux obtenus avec la méthode SWIP. Pour les résultats présentés au chapitre 2, nous avons collaboré avec Paolo Zunino[4]. Les tests numériques, effectués

[4]MOX, Dipartimento di Matematica 'F. Brioschi', Politecnico di Milano

par Paolo Zunino, ont été réalisés avec le logiciel gratuit freeFEM++ [5], développé par Frédérich Hecht, Antoine Le Hyaric et Olivier Pironneau au Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie. Le chapitre 2 est le sujet d'un article soumis pour publication dans IMA Journal of Numerical Analysis [51].

Dans le chapitre 3 nous présentons un premier estimateur d'erreur par résidus. Un résultat sous forme abstraite montre d'abord comment l'erreur dans la norme d'énergie peut être contrôlée par l'erreur de non-conformité et deux autres termes. Le premier de ces termes fait intervenir la forme bilinéaire associée au problème et l'erreur d'approximation, et le deuxieme dépend de la partie antisymétrique de cette forme bilinéaire et l'erreur de non-conformité. Nous considerons d'abord l'équation de Poisson, où la partie antisymétrique de la forme bilinéaire est nulle. Nous obtenons un estimateur où toutes les constantes sont calculables. L'indicateur d'erreur est divisé en trois parties : une qui mesure la non-conformité de la solution calculée, une qui dépend du résidu et une qui mesure la non-conformité des flux. Ensuite nous considérons l'équation d'advection-diffusion-réaction à laquelle nous appliquons la même procédure. Nous nous sommes inspirés du travail de Verfürth [98] pour définir la robustesse de l'estimateur en régime de nombre de Péclet élevé. Celui-ci est dit robuste si les constantes intervenant dans les estimateurs locaux dépendent du nombre de Péclet sous la forme $C_1 + C_2 \min(\mathrm{Pe}, \rho)$, où ρ dépend du champ advectif et du tenseur de diffusion, mais pas du maillage. L'indicateur d'erreur est composé de quatre parties : une qui dépend du résidu, une qui mesure la non-conformité des flux (identique à celle trouvée pour l'équation de Poisson) et deux qui mesurent la non-conformité de la solution calculée. Une amélioration particulière à mentionner est dans le résidu, où nous considérons le résidu auquel sa projection L^2 sur l'élément est soustraite, ce qui augmente considérablement l'efficacité de l'estimateur. Nous avons utilisé pour cela le fait que les fonctions constantes par morceaux sont dans l'espace d'approximation des méthodes GD. En ce qui concerne l'efficacité locale des indicateurs, notons que seulement l'efficacité de l'erreur de non-conformité, qui est évaluée en utilisant l'interpolé de Oswald, dépend de l'hétérogénéité du tenseur de diffusion. La robustesse de tous les autres indicateurs est une conséquence des propriétés des poids de la méthode SWIP. Les tests numériques présentés à la fin du chapitre montrent la bonne convergence de l'estimateur et une efficacité en cohérence avec la notion de robustesse de Verfürth. Tous les tests numériques ont été réalisés avec un code écrit en C++, dont le noyau GD a été développé au Cermics par Daniele

[5] http ://www.freefem.org

Di Pietro[6]. Les maillages structurés ont été construits avec le logiciel gratuit gmsh[7] développé par Christophe Geuzaine et Jean-François Remacle. Les maillages non-structurés et l'adaptation de maillage basée sur les indicateurs d'erreur ont été réalisés avec le logiciel Matlab[8]. Le travail présenté au chapitre 3 est soumis pour publication dans Journal of Computational Mathematics [49].

Dans le chapitre 4 nous présentons un deuxième estimateur d'erreur par résidus, utilisant cette fois la construction d'un champ vectoriel auxiliaire. Nous donnons d'abord un estimateur abstrait pour l'équation de Poisson. Cet estimateur est quasi-optimal, c'est-à-dire que l'indice d'efficacité est égal à $\sqrt{2}$. La norme d'énergie est ici contrôlée par deux termes : le minimum de l'erreur de non-conformité, considérant toutes les fonctions possibles de $H_0^1(\Omega)$, et un autre terme où le minimum est pris en considérant toutes les fonctions de $H(\mathrm{div}, \Omega)$. Pour pouvoir calculer l'estimateur, nous avons choisi à nouveau d'utiliser l'interpolé de Oswald et d'utiliser les espaces de fonctions vectorielles de Raviart-Thomas-Nédélec ($\subset H(\mathrm{div}, \Omega)$). En considérant une solution approchée affine par morceaux ($p = 1$), la fonction de $H(\mathrm{div}, \Omega)$ peut être construite grâce à la résolution de problèmes locaux à 3 ou 8 degrés de liberté, correspondant respectivement aux degrés de liberté des éléments finis RT_0 et RT_1. La convergence du résidu est d'un ordre plus élevé en utilisant une reconstruction des flux dans l'espace RT_1. L'estimateur ainsi obtenu ne dépend pas de la régularité du maillage, et ne nécessite pas de terme additionnel pour traiter les oscillations du terme source. Passant ensuite à l'équation d'advection-diffusion-réaction, nous construisons un deuxième champ vectoriel basé cette fois sur les flux advectifs. En utilisant une construction de degré maximal, nous montrons que sous des hypothèses minimales sur le coefficient de réaction et la divergence du champ advectif, le terme de résidu est de la forme $\|f - \pi_p f\|_{0,T}$. Dans les tests numériques, nous examinons d'abord le cas de la diffusion pure, et en particulier nous observons que le résidu converge bien à l'ordre prévu par la théorie, que ce soit pour le cas d'une construction dans l'espace RT_0 ou dans l'espace RT_1. De plus, l'indice d'efficacité du nouvel estimateur est très proche de 1. Nous présentons également une comparaison avec deux autres estimateurs, notamment celui obtenu au chapitre précédent. Même les cas tests avec une solution exacte qui présente une singularité dans le domaine de calcul due aux hétérogénéités de la diffusion montrent la bonne convergence de l'estimateur et un très bon indice d'efficacité. Les cas tests ont été réalisés sur des maillages structurés et non-structurés. Pour le cas avec advection domi-

[6] Institut Français du Pétrole
[7] www.geuz.org/gmsh/
[8] www.mathworks.com

nante, les estimateurs montrent toujours une bonne convergence, l'indice d'efficacité étant en cohérence avec la notion de robustesse de Verfürth. Pour tous les cas tests (réalisés avec $p = 1$ dans la méthode SWIP), l'indice d'efficacité ne change pas beaucoup en passant d'une construction dans l'espace RT_0 à une construction dans l'espace RT_1. Enfin nous présentons des maillages adaptifs basés sur l'indicateur d'erreur avec une construction dans l'espace RT_0 et pour un problème de diffusion pure avec singularité. Conformément aux attentes, le maillage devient plus raffiné là où la solution exacte présente une singularité. Pour les résultats présentés au chapitre 4 nous avons collaboré avec Martin Vohralík[9], qui a été à l'origine d'une grande partie de la rédaction. Les maillages structurés ont été construits avec le logiciel gratuit gmsh. Les maillages non-structurés et l'adaptation de maillage basée sur les indicateurs d'erreur ont été réalisés avec le logiciel Matlab. Tous les tests numériques ont été réalisés par l'auteur de la thèse en utilisant le code C++ précédemment mentionné. Le travail présenté au chapitre 4 a été soumis pour publication dans SIAM Journal on Numerical Analysis [50].

Le chapitre 5 dresse la conclusion de ce travail de thèse et propose quelques perspectives.

[9]Laboratoire Jacques-Louis Lions, Université Pierre et Matie Curie (Paris 6)

Chapitre 2

A Discontinuous Galerkin method with weighted averages

Submitted to IMA Journal of Numerical Analysis under the title 'A Discontinuous Galerkin method with weighted averages for advection-diffusion-reaction equations with locally small and anisotropic diffusivity'.

Alexandre Ern[1], Annette F. Stephansen[1,2] and Paolo Zunino[3]

Abstract: We propose and analyze a symmetric weighted interior penalty (SWIP) method to approximate in a Discontinuous Galerkin framework advection-diffusion-reaction equations with anisotropic and discontinuous diffusivity. The originality of the method consists in the use of diffusivity-dependent weighted averages to better cope with locally small diffusivity (or equivalently with locally high Péclet numbers) on fitted meshes. The analysis yields convergence results for the natural energy norm that are optimal with respect to mesh-size and robust with respect to diffusivity. The convergence results for the advective derivative are optimal with respect to mesh-size and robust for isotropic diffusivity, as well as for anisotropic diffusivity if the cell Péclet numbers evaluated with the largest eigenvalue of the diffusivity tensor are large enough. Numerical results are presented to illustrate the performance of the proposed scheme.

[1]Cermics, Ecole des Ponts, ParisTech, 6 et 8 avenue Blaise Pascal, Champs sur Marne, 77455 Marne la Vallée Cedex 2, France.

[2]Andra, Parc de la Croix-Blanche, 1-7 rue Jean Monnet, 92298 Châtenay-Malabry cedex, France.

[3]MOX, Dipartimento di Matematica "F. Brioschi", Politecnico di Milano, via Bonardi 9, 20133 Milano, Italy.

2.1 Introduction

Since their introduction over thirty years ago [69,83], Discontinuous Galerkin (DG) methods have emerged as an attractive tool to approximate numerous PDEs in the engineering sciences. Here we are primarily interested in advection-diffusion-reaction equations with anisotropic (e.g., tensor-valued) and heterogeneous (e.g., non-smooth) diffusivity. Such equations are encountered, for instance, in groundwater flow models which constitute the motivation for the present work.

The analysis of DG methods to approximate advection-diffusion-reaction equations is extensively covered in [59]. This work already addresses anisotropic and heterogeneous diffusivity. However, one particular aspect that deserves further attention is that where the diffusivity becomes very small in *some* parts of the computational domain. Indeed, in this case it is well-known that the presence of an advective field can trigger internal layers. In the locally vanishing diffusivity limit, the solution becomes discontinuous on the interfaces where the advective field flows from the vanishing-diffusivity region towards the nonvanishing-diffusivity region. This situation has been analyzed in [52] and, more recently, in [34,38]. For (very) small but positive diffusivity, the usual DG methods meet with difficulties in the presence of internal layers that are not sufficiently resolved by the mesh. Indeed, these methods are designed to weakly enforce continuity of the discrete solution across mesh interfaces, but because internal layers are under-resolved, the exact solution is better approximated by a discontinuous function at the interfaces adjacent to internal layers. One possible remedy is to consider a hard-wired modification of the DG method at those interfaces, as already proposed in [59] and, more recently, in [48]. However, a more satisfactory approach would be to design a DG method that can handle internal layers in an automated fashion. This is the purpose of the present work. The key ingredient is the use of weighted instead of arithmetic averages in certain interface terms of the DG method, with weights depending on the diffusivity on both sides of the interface. The present method relies on the (mild) assumption that fitted meshes are used, i.e., that discontinuities in the diffusivity are aligned with the mesh. When this assumption is not possible (e.g., in the case of nonlinear diffusivity), the present method is not expected to behave better than the usual DG methods, since all methods will suffer from the fact that they attempt to approximate a rough solution within some mesh elements.

The idea of utilizing weighted averages stems from the mortar finite-element method originally proposed by Nitsche [76, 77]. This method imposes weakly the continuity of fluxes between different regions. Various authors have highlighted the possibility of using

an average with weights that differ from one half; see [54–56, 94] where several mortaring techniques are presented to match conforming finite elements on possibly nonconforming computational meshes. In the cited works, weighted averages are introduced as a generalization of standard averages and the analysis is carried out in the general framework, but a possible dependency of the weights on the coefficients of the problem is not considered. This dependency was investigated recently in [25] for isotropic advection-diffusion-reaction problems, using a weighted interior penalty technique with mortars; when applied elementwise, this approach yields a DG method. It was shown in [25] that a specific choice of weights improves the stability of the scheme when the diffusivity takes locally small values. The reason why weighted averages are needed to properly handle internal layers is rooted in the dissipative structure of the underlying Friedrichs's system. The design of the corresponding DG bilinear form, where dissipation at the discrete level is enforced by a consistency term involving averages, has been recently proposed in [43]. The extension to advection-diffusion-reaction equations including the locally vanishing diffusivity limit is analyzed in [38].

In the present work, we extend the DG method implicitly derived in [25] for isotropic diffusivity to anisotropic problems. This task is not as simple as it may appear on first sight since the presence of internal layers now depends on the spectral structure of the diffusivity tensor on both sides of each mesh interface. The spectral structure also raises the question of the appropriate choice of the penalty term in the DG method at each mesh interface. The analysis presented below will tackle these issues.

We design and analyze one specific DG method with weighted averages, namely the Symmetric Weighted Interior Penalty (SWIP) method, obtained by modifying the well-known (Symmetric) Interior Penalty (IP) method [10, 15]. Many other well-known DG methods, including the Local Discontinuous Galerkin method [33] and the Nonsymmetric Interior Penalty Galerkin method [89], can also be modified to fit the present scope; for brevity, these developments are omitted herein.

This paper is organized as follows: Section 2.2 presents the setting under scrutiny and formulates the SWIP method, while Section 2.3 contains the error analysis in the natural energy norm for the problem. The estimate is fully robust, meaning that the constant in the error upper bound is independent of both heterogeneities and anisotropies in the diffusivity. Section 2.4 is concerned with the error analysis on the advective derivative. The derived estimate is again robust with respect to heterogeneities in the diffusivity, but the constant in the error upper bound can in some cases depend on local anisotropies. Robustness is achieved for instance if the cell Péclet numbers evaluated with the largest eigenvalue of the

diffusivity tensor are large enough. Numerical results, including comparisons with the more usual IP methods, are presented in Section 2.5 and illustrate the benefits of using weighted interior penalties to approximate advection-diffusion-reaction equations with locally small and anisotropic diffusivity. Finally, Section 2.6 contains some concluding remarks.

2.2 The SWIP method

Let Ω be a domain in $\mathbb{R}^d$ with boundary $\partial\Omega$ in space dimension $d \in \{2,3\}$. We consider the following advection-diffusion-reaction equation with homogeneous Dirichlet boundary conditions:

$$\begin{cases} -\nabla\cdot(K\nabla u) + \beta\cdot\nabla u + \mu u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{2.1}$$

Here $\mu \in L^\infty(\Omega)$, $\beta \in [W^{1,\infty}(\Omega)]^d$, the diffusivity tensor K is a symmetric, positive definite field in $[L^\infty(\Omega)]^{d,d}$ and $f \in L^2(\Omega)$. The regularity assumption on β can be relaxed, but is sufficient for the present purpose. The weak formulation of (2.1) consists of finding $u \in H_0^1(\Omega)$ such that

$$(K\nabla u, \nabla v)_{0,\Omega} + (\beta\cdot\nabla u, v)_{0,\Omega} + (\mu u, v)_{0,\Omega} = (f, v)_{0,\Omega} \quad \forall v \in H_0^1(\Omega) \tag{2.2}$$

where $(\cdot,\cdot)_{0,\Omega}$ denotes the L^2-scalar product on Ω. Henceforth, we assume that

$$\mu - \tfrac{1}{2}\nabla\cdot\beta \geq \mu_0 > 0 \qquad \text{a.e in } \Omega. \tag{2.3}$$

Furthermore, we assume that the smallest eigenvalue of K is bounded from below by a positive (but possibly very small) constant. Then, owing to the Lax–Milgram Lemma, (2.2) is well–posed.

Let $\{\mathcal{T}_h\}_{h>0}$ be a shape-regular family of affine triangulations of the domain Ω. The meshes $\mathcal{T}_h$ may possess hanging nodes. For simplicity we assume that the meshes cover Ω exactly, i.e., Ω is a polyhedron. A generic element in $\mathcal{T}_h$ is denoted by T, h_T denotes the diameter of T and n_T its outward unit normal. Set $h = \max_{T\in\mathcal{T}_h} h_T$. We assume without loss of generality that $h \leq 1$. Let $p \geq 1$. We define the classical DG approximation space

$$V_h = \{v_h \in L^2(\Omega); \forall T \in \mathcal{T}_h, v_h|_T \in \mathbb{P}_p\}, \tag{2.4}$$

where $\mathbb{P}_p$ is the set of polynomials of total degree less than or equal to p. Henceforth, we assume that the discontinuities in the diffusivity tensor are aligned with the mesh. This is a

mild assumption in the context of linear problems. Moreover, for the sake of simplicity, we assume that the diffusivity tensor K is piecewise constant on $\mathcal{T}_h$. This assumption, which is reasonable in the context of groundwater flow models, can be generalized by assuming a smooth enough behavior of K inside each mesh element.

We say that F is an interior face of the mesh if there are $T^-(F)$ and $T^+(F)$ in $\mathcal{T}_h$ such that $F = T^-(F) \cap T^+(F)$. We set $\mathcal{T}(F) = \{T^-(F), T^+(F)\}$ and let n_F be the unit normal vector to F pointing from $T^-(F)$ towards $T^+(F)$. The analysis hereafter does not depend on the arbitrariness of this choice. Similarly, we say that F is a boundary face of the mesh if there is $T(F) \in \mathcal{T}_h$ such that $F = T(F) \cap \partial\Omega$. We set $\mathcal{T}(F) = \{T(F)\}$ and let n_F coincide with the outward normal to $\partial\Omega$. All the interior (resp., boundary) faces of the mesh are collected into the set $\mathcal{F}_h^i$ (resp., $\mathcal{F}_h^{\partial\Omega}$) and we let $\mathcal{F}_h = \mathcal{F}_h^i \cup \mathcal{F}_h^{\partial\Omega}$. Henceforth, we shall often deal with functions that are double-valued on $\mathcal{F}_h^i$ and single-valued on $\mathcal{F}_h^{\partial\Omega}$. This is the case, for instance, of functions in V_h. On interior faces, when the two branches of the function in question, say v, are associated with restrictions to the neighboring elements $T^\mp(F)$, these branches are denoted by $v^\mp$ and the jump of v across F is defined as

$$[\![v]\!]_F = v^- - v^+. \tag{2.5}$$

On a boundary face $F \in \mathcal{F}^{\partial\Omega}$, we set $[\![v]\!]_F = v|_F$. Furthermore, on an interior face $F \in \mathcal{F}_h^i$, we define the standard (arithmetic) average as $\{v\}_F = \frac{1}{2}(v^- + v^+)$. For convenience, we set $\{v\}_F = \frac{1}{2}v|_F$ on $F \in \mathcal{F}^{\partial\Omega}$. The subscript F in the above jumps and averages is omitted if there is no ambiguity.

The L^2-scalar product and its associated norm on a subset $R \subset \Omega$ (evaluated with the appropriate Lebesgue's measure) are indicated by the subscript $0, R$. For $s \geq 1$, a norm (seminorm) with the subscript s, R designates the usual norm (seminorm) in $H^s(R)$. When the region R is the boundary of a mesh element ∂T and the arguments in the scalar product or the norm are double-valued functions, it is implicitly assumed that the value considered is that of the branch associated with the restriction to T. For $s \geq 1$, $H^s(\mathcal{T}_h)$ denotes the usual broken Sobolev space on $\mathcal{T}_h$ and for $v \in H^1(\mathcal{T}_h)$, $\nabla_h v$ denotes the piecewise gradient of v, that is, $\nabla_h v \in [L^2(\Omega)]^d$ and for all $T \in \mathcal{T}_h$, $(\nabla_h v)|_T = \nabla(v|_T)$. It is also convenient to set $V(h) = H^2(\mathcal{T}_h) + V_h$.

The formulation of the SWIP method requires two parameters. As in the formulation of the usual IP method we introduce a scalar- and single-valued function γ_F defined on $\mathcal{F}_h$. The purpose of this function is to penalize jumps across interior faces and values at boundary faces. Additionally, we define a scalar- and double-valued function $\omega_{T,F}$ for $T \in \mathcal{T}_h$ and $F \subset \partial T$, $F \in \mathcal{F}_h^i$. This function, which is not present in the usual IP method,

is used to evaluate weighted averages of diffusive fluxes. On an interior face $F \in \mathcal{F}_h^i$, the values taken by the two branches of $\omega_{T,F}$ are denoted by $\omega_{T^{\mp}(F),F}$. Henceforth, it is assumed that for all $F \in \mathcal{F}_h^i$, both values are non-negative and that

$$\omega_{T^-(F),F} + \omega_{T^+(F),F} = 1. \tag{2.6}$$

For $v \in V(h)$, we define the weighted average of the diffusive flux $K\nabla_h v$ on an interior face $F \in \mathcal{F}_h^i$ as

$$\{K\nabla_h v\}_\omega = \omega_{T^-(F),F}(K\nabla_h v)^- + \omega_{T^+(F),F}(K\nabla_h v)^+. \tag{2.7}$$

For convenience, we extend the above definitions to boundary faces as follows: on $F \in \mathcal{F}_h^{\partial\Omega}$, $\omega_{T,F}$ is single-valued and equal to 1, and we set $\{K\nabla v\}_\omega = K\nabla v$.

The SWIP bilinear form $B_h(\cdot,\cdot)$ is defined on $V(h) \times V(h)$ as follows

$$\begin{aligned} B_h(v,w) &= (K\nabla_h v, \nabla_h w)_{0,\Omega} + ((\mu - \nabla\cdot\beta)v, w)_{0,\Omega} - (v, \beta\cdot\nabla_h w)_{0,\Omega} \\ &+ \sum_{F\in\mathcal{F}_h} \big((\gamma_F[\![v]\!],[\![w]\!])_{0,F} - (n_F^t\{K\nabla_h v\}_\omega,[\![w]\!])_{0,F} - (n_F^t\{K\nabla_h w\}_\omega,[\![v]\!])_{0,F}\big) \\ &+ \sum_{F\in\mathcal{F}_h} (\beta\cdot n_F\{v\},[\![w]\!])_{0,F}. \end{aligned} \tag{2.8}$$

The SWIP bilinear form can equivalently be expressed, after integrating the advective derivative by parts, as

$$\begin{aligned} B_h(v,w) &= (K\nabla_h v, \nabla_h w)_{0,\Omega} + (\mu v, w)_{0,\Omega} + (\beta\cdot\nabla_h v, w)_{0,\Omega} \\ &+ \sum_{F\in\mathcal{F}_h} \big((\gamma_F[\![v]\!],[\![w]\!])_{0,F} - (n_F^t\{K\nabla_h v\}_\omega,[\![w]\!])_{0,F} - (n_F^t\{K\nabla_h w\}_\omega,[\![v]\!])_{0,F}\big) \\ &- \sum_{F\in\mathcal{F}_h} (\beta\cdot n_F\{w\},[\![v]\!])_{0,F}. \end{aligned} \tag{2.9}$$

Both (2.8) and (2.9) will be used in the analysis. The discrete problem consists of finding $u_h \in V_h$ such that

$$B_h(u_h, v_h) = (f, v_h)_{0,\Omega} \qquad \forall v_h \in V_h. \tag{2.10}$$

The penalty parameter γ_F is defined as

$$\forall F \in \mathcal{F}_h, \qquad \gamma_F = \alpha\frac{\gamma_{K,F}}{h_F} + \gamma_{\beta,F}, \tag{2.11}$$

where α is a positive scalar (α can also vary from face to face) and where

$$\forall F \in \mathcal{F}_h^i, \qquad \gamma_{K,F} = (\omega_{T^-(F),F})^2 \delta_{K,F-} + (\omega_{T^+(F),F})^2 \delta_{K,F+} \tag{2.12}$$

$$\forall F \in \mathcal{F}_h^{\partial\Omega}, \qquad \gamma_{K,F} = \delta_{K,F}, \tag{2.13}$$

$$\forall F \in \mathcal{F}_h, \qquad \gamma_{\beta,F} = \tfrac{1}{2}|\beta \cdot n_F|, \tag{2.14}$$

with $\delta_{K,F\mp} = n_F^t K^{\mp} n_F$ if $F \in \mathcal{F}_h^i$ and $\delta_{K,F} = n_F^t K n_F$ if $F \in \mathcal{F}_h^{\partial\Omega}$. Note that the choice for $\gamma_{\beta,F}$ amounts to the usual upwind scheme to stabilize the advective derivative. As for any symmetric IP method, the size of the penalty parameter α is assumed to be large enough. This assumption is made for the rest of this work. The minimal value for α depends on the actual value of the constant arising in the trace inequality (2.17) stated below; it can be determined from the proof of Lemma 2.1 to ensure coercivity. Because they are standard, these developments are omitted.

For the error analysis in the energy norm (see Section 2.3), no other assumption than (2.6) is made for the weights. In particular, it is possible to choose $\omega_{T^{\mp}(F),F} = \frac{1}{2}$, in which case the SWIP bilinear form B_h reduces to the standard IP bilinear form with the penalty parameter scaling as the standard average of the diffusivity in the normal direction; this method has been analyzed in [53]. Note also that the choice made in [59] for the penalty parameter is different since it involves the maximum eigenvalue of K.

For the error analysis in the advective derivative (see Section 2.4), a specific choice of the weights differing from $\omega_{T^{\mp}(F),F} = \frac{1}{2}$ has to be made to yield robust error estimates with respect to the diffusivity. Specifically, we shall set

$$\omega_{T^-(F),F} = \frac{\delta_{K,F+}}{\delta_{K,F+} + \delta_{K,F-}}, \qquad \omega_{T^+(F),F} = \frac{\delta_{K,F-}}{\delta_{K,F+} + \delta_{K,F-}}, \tag{2.15}$$

and thus

$$\forall F \in \mathcal{F}_h^i, \qquad \gamma_{K,F} = \frac{\delta_{K,F+}\delta_{K,F-}}{\delta_{K,F+} + \delta_{K,F-}}. \tag{2.16}$$

Note that with this choice $\gamma_{K,F} = \omega_{T^-(F),F}\delta_{K,F-} = \omega_{T^+(F),F}\delta_{K,F+}$, and that $2\gamma_{K,F}$ is the harmonic average of the normal component of the diffusivity tensor across the interface. Observe also that $\gamma_{K,F} \le \inf(\delta_{K,F-}, \delta_{K,F+})$, a point that becomes important to ensure even the consistency of the method when the diffusivity is actually allowed to vanish locally, see [38]. The numerical results presented in Section 2.5 show that also in the energy norm, the DG method behaves better if the weights are chosen according to (2.15). Hence, we recommend this choice whenever the diffusivity exhibits heterogeneities.

2.3 Error analysis in the energy norm

The goal of this section is to establish an error estimate for the SWIP method in the energy norm, the estimate being robust with respect to heterogeneities and anisotropies in the diffusivity. The analysis is performed using fairly standard arguments, i.e., by establishing coercivity, consistency and continuity properties for the SWIP bilinear form in the spirit of Strang's Second Lemma [42].

In the sequel, the symbol $\lesssim$ indicates an inequality involving a positive constant C independent of the mesh family and of the diffusivity. The constant C can depend on $\|\beta\|_{[W^{1,\infty}(\Omega)]^d}$, $\|\mu\|_{L^\infty(\Omega)}$, μ_0^{-1} (see (2.3)), and the shape-regularity of the mesh family. Without loss of generality, it can be assumed that the problem data is normalized so that $\|\beta\|_{[W^{1,\infty}(\Omega)]^d}$ is of order unity. We will not be concerned with the dependency on $\|\mu\|_{L^\infty(\Omega)}$ since we are not interested in strong reaction regimes. The dependency on μ_0^{-1} can be addressed by means of Poincaré inequalities; this will not be further discussed here. Owing to the shape-regularity of the mesh family, the following inverse trace and inverse inequalities hold: For all $T \in \mathcal{T}_h$ and for all $v_h \in V_h$,

$$\|v_h\|_{0,\partial T} \lesssim h_T^{-\frac{1}{2}} \|v_h\|_{0,T}, \tag{2.17}$$

$$\|\nabla_h v_h\|_{0,T} \lesssim h_T^{-1} \|v_h\|_{0,T}, \tag{2.18}$$

which result from the shape regularity of the mesh family $\{\mathcal{T}_h\}_{h>0}$.

For a function $v \in V(h)$, we consider the following jump seminorms

$$|[\![v]\!]|_\sigma^2 = \sum_{F \in \mathcal{F}_h} |[\![v]\!]|_{\sigma,F}^2, \qquad |[\![v]\!]|_{\sigma,F}^2 = (\sigma[\![v]\!], [\![v]\!])_{0,F}, \tag{2.19}$$

with $\sigma := \gamma_{\beta,F}$, $\sigma := \gamma_{K,F}$ or $\sigma := \gamma_F$. The natural energy norm with which to equip $V(h)$ is

$$\|v\|_{h,B} = \|v\|_{0,\Omega} + \|\kappa \nabla_h v\|_{0,\Omega} + |[\![v]\!]|_{\gamma_F} \tag{2.20}$$

where κ denotes the (unique) symmetric positive definite tensor-valued field such that $\kappa^2 = K$ a.e. in Ω.

Lemma 2.1. (Coercivity) *The bilinear form B_h is $\|\cdot\|_{h,B}$-coercive, i.e., for all $v_h \in V_h$,*

$$B_h(v_h, v_h) \gtrsim \|v_h\|_{h,B}^2. \tag{2.21}$$

Proof. Let $v_h \in V_h$. Taking $v = w = v_h$ in (2.8) yields

$$B_h(v_h, v_h) = \|\kappa\nabla_h v_h\|_{0,\Omega}^2 + (\mu v_h, v_h)_{0,\Omega} - ((\nabla\cdot\beta)v_h, v_h)_{0,\Omega} - (v_h, \beta\cdot\nabla_h v_h)_{0,\Omega} + |[\![v_h]\!]|_{\gamma_F}^2 - \sum_{F\in\mathcal{F}_h} 2(n_F^t\{K\nabla v_h\}_\omega, [\![v_h]\!])_{0,F} + \sum_{F\in\mathcal{F}_h} (\beta\cdot n_F\{v_h\}, [\![v_h]\!])_{0,F}. \tag{2.22}$$

Integrating by parts the fourth term on the right hand side of (2.22) and owing to hypothesis (2.3), we obtain

$$(\mu v_h, v_h)_{0,\Omega} - ((\nabla\cdot\beta)v_h, v_h)_{0,\Omega} - (v_h, \beta\cdot\nabla_h v_h)_{0,\Omega} + \sum_{F\in\mathcal{F}_h} (\beta\cdot n_F\{v_h\}, [\![v_h]\!])_{0,F} = ((\mu - \tfrac{1}{2}\nabla\cdot\beta)v_h, v_h)_{0,\Omega} \gtrsim \|v_h\|_{0,\Omega}^2. \tag{2.23}$$

Consider now the sixth term in the right-hand side of (2.22). Let $F \in \mathcal{F}_h$. First, observe that owing to Young's inequality

$$|2(n_F^t\omega_{T^\mp(F),F}(K\nabla_h v_h)^\mp, [\![v_h]\!])_{0,F}| = |2((\kappa\nabla_h v_h)^\mp, \omega_{T^\mp(F),F}\kappa^\mp n_F[\![v_h]\!])_{0,F}| \le h_F\alpha_0\|(\kappa\nabla_h v_h)^\mp\|_{0,F}^2 + \frac{1}{\alpha_0}\left(\frac{(\omega_{T^\mp(F),F})^2\delta_{K,F^\mp}}{h_F}[\![v_h]\!], [\![v_h]\!]\right)_{0,F},$$

where $\alpha_0 > 0$ can be chosen as small as needed. Using the trace inverse inequality (2.17) and the definition of $\gamma_{K,F}$ (2.12)-(2.13) yields

$$|2(n_F^t\{K\nabla_h v_h\}_\omega, [\![v_h]\!])_{0,F}| \lesssim \alpha_0\|\kappa\nabla_h v_h\|_{0,\mathcal{T}(F)}^2 + \frac{1}{\alpha_0 h_F}|[\![v_h]\!]|_{\gamma_{K,F}}^2.$$

The end of the proof is classical since α in (2.11) can be chosen to be large enough. □

Lemma 2.2. (Consistency) *Let u solve (2.2) and let u_h solve (2.10). Assume that $u \in H^2(\mathcal{T}_h)$. Then*

$$\forall v_h \in V_h, \qquad B_h(u - u_h, v_h) = 0 \tag{2.24}$$

Proof. Let $v_h \in V_h$. Since $u \in H_0^1(\Omega)$, (2.9) yields

$$B_h(u, v_h) = (K\nabla u, \nabla_h v_h)_{0,\Omega} + (\mu u, v_h)_{0,\Omega} + (\beta\cdot\nabla u, v_h)_{0,\Omega} - \sum_{F\in\mathcal{F}_h} (n_F^t\{K\nabla u\}_\omega, [\![v_h]\!])_{0,F}.$$

Using the fact that $n_F^t K\nabla u$ is continuous on interior faces yields $n_F^t\{K\nabla u\}_\omega = (\omega_{T^-(F),F} + \omega_{T^+(F),F})n_F^t K\nabla u = n_F^t K\nabla u$ owing to (2.6). Hence, integrating by parts leads to

$$(K\nabla u, \nabla_h v_h)_{0,\Omega} - \sum_{F\in\mathcal{F}_h} (n_F^t\{K\nabla u\}_\omega, [\![v_h]\!])_{0,F} = -\sum_{T\in\mathcal{T}_h} (\nabla\cdot(K\nabla u), v_h)_{0,T}.$$

As a result,

$$B_h(u, v_h) = \sum_{T\in\mathcal{T}_h} (-\nabla\cdot(K\nabla u) + \beta\cdot\nabla u + \mu u, v_h)_{0,T} = (f, v_h)_{0,\Omega} = B_h(u_h, v_h),$$

yielding (2.24). □

We now establish a continuity property for the SWIP bilinear form B_h. To this purpose, we introduce on $V(h)$ the norm

$$\|v\|_{h,\frac{1}{2}} = \|v\|_{h,B} + \left(\sum_{T\in\mathcal{T}_h} \|v\|_{0,\partial T}^2\right)^{\frac{1}{2}} + \left(\sum_{T\in\mathcal{T}_h} h_T\|\kappa\nabla_h v\|_{0,\partial T}^2\right)^{\frac{1}{2}}. \tag{2.25}$$

Let $V_h^\perp = \{v \in V(h), \forall v_h \in V_h, (v, v_h)_{0,\Omega} = 0\}$.

Lemma 2.3. (Continuity) *The following holds:*

$$\forall (v, w_h) \in V_h^\perp \times V_h, \qquad |B_h(v, w_h)| \lesssim \|v\|_{h,\frac{1}{2}}\|w_h\|_{h,B}. \tag{2.26}$$

Proof. Let $(v, w_h) \in V_h^\perp \times V_h$. The first two terms in (2.8) are easily bounded as

$$|(K\nabla_h v, \nabla_h w_h)_{0,\Omega}| + |((\mu - \nabla\cdot\beta)v, w_h)_{0,\Omega}| \lesssim \|v\|_{h,B}\|w_h\|_{h,B}.$$

To bound the third term, let $\overline{\beta}$ be the piecewise constant, vector-valued field equal to the mean value of β on each $T \in \mathcal{T}_h$. Then,

$$\begin{aligned}(v, \beta\cdot\nabla_h w_h)_{0,\Omega} &= (v, \overline{\beta}\cdot\nabla_h w_h)_{0,\Omega} + (v, (\beta - \overline{\beta})\cdot\nabla_h w_h)_{0,\Omega}\\ &= (v, (\beta - \overline{\beta})\cdot\nabla_h w_h)_{0,\Omega},\end{aligned}$$

since $\overline{\beta}\cdot\nabla_h w_h \in V_h$ and $v \in V_h^\perp$. Moreover, since $\beta \in [W^{1,\infty}(\Omega)]^d$,

$$\forall T \in \mathcal{T}_h, \qquad \|\beta - \overline{\beta}\|_{[L^\infty(T)]^d} \lesssim h_T,$$

so that the inverse inequality (2.18) yields

$$|(v, \beta\cdot\nabla_h w_h)_{0,\Omega}| \lesssim \|v\|_{0,\Omega}\|w_h\|_{0,\Omega} \le \|v\|_{h,B}\|w_h\|_{h,B}.$$

Furthermore, proceeding as in the proof of Lemma 2.1 yields, for all $F \in \mathcal{F}_h$,

$$|(n_F^t\{K\nabla_h v\}_\omega, [\![w_h]\!])_{0,F}| \lesssim \left(\sum_{T\in\mathcal{T}(F)} h_T^{\frac{1}{2}}\|\kappa\nabla_h v\|_{0,\partial T}\right) h_F^{-\frac{1}{2}}|[\![w_h]\!]|_{\gamma_{K,F}}$$

and

$$|(n_F^t\{K\nabla_h w_h\}_\omega, [\![v]\!])_{0,F}| \lesssim h_F^{-\frac{1}{2}} |[\![v]\!]|_{\gamma_{K,F}} \|\kappa\nabla_h w_h\|_{0,\mathcal{T}(F)},$$

so that

$$\sum_{F\in\mathcal{F}_h} \left(|(n_F^t\{K\nabla v\}_\omega, [\![w_h]\!])_{0,F}| + |(n_F^t\{K\nabla w_h\}_\omega, [\![v]\!])_{0,F}| \right) \lesssim \|v\|_{h,\frac{1}{2}} \|w_h\|_{h,B}.$$

For the remaining terms, we obtain

$$\begin{aligned}\sum_{F\in\mathcal{F}_h} |(\gamma_F [\![v]\!], [\![w_h]\!])_{0,F}| + \sum_{F\in\mathcal{F}_h} |(\beta\cdot n_F\{v\}, [\![w_h]\!])_{0,F}| \\ \lesssim |[\![v]\!]|_{\gamma_F} |[\![w_h]\!]|_{\gamma_F} + \sum_{F\in\mathcal{F}_h} \|\{v\}\|_{0,F} |[\![w_h]\!]|_{\gamma_{\beta,F}} \leq \|v\|_{h,\frac{1}{2}} \|w_h\|_{h,B}.\end{aligned}$$

This completes the proof since $\|\cdot\|_{h,B} \leq \|\cdot\|_{h,\frac{1}{2}}$. □

Theorem 2.4. *Let $\Pi_h u$ be the L^2-projection of u onto V_h. Then,*

$$\|u - u_h\|_{h,B} \lesssim \|u - \Pi_h u\|_{h,\frac{1}{2}}. \tag{2.27}$$

Proof. Owing to Lemmata 2.1, 2.2 and 2.3,

$$\begin{aligned}\|u_h - \Pi_h u\|_{h,B} &\lesssim \frac{B_h(u_h - \Pi_h u, u_h - \Pi_h u)}{\|u_h - \Pi_h u\|_{h,B}} = \frac{B_h(u - \Pi_h u, u_h - \Pi_h u)}{\|u_h - \Pi_h u\|_{h,B}} \\ &\lesssim \|u - \Pi_h u\|_{h,\frac{1}{2}}.\end{aligned} \tag{2.28}$$

We complete the proof by applying the triangle inequality and using the fact that $\|\cdot\|_{h,B} \leq \|\cdot\|_{h,\frac{1}{2}}$. □

Remark. Estimate (2.27) yields an error upper bound in the natural energy norm with a constant independent of the diffusivity tensor. Furthermore, if the exact solution is smooth enough locally on each mesh cell, namely $u \in H^{p+1}(\mathcal{T}_h)$, it is readily seen using standard approximation properties for the L^2-orthogonal projector Π_h, that the upper bound converges as h^p, which is optimal.

We now prove that under some assumptions, the error estimate in the L^2-norm can be improved using the Aubin-Nitsche duality argument. Let $\lambda_{m,K}$ denote the lowest eigenvalue of K in Ω and set $\lambda_{M,K} = \max(1, \lambda_K)$ where λ_K denotes the largest eigenvalue of K in Ω. We introduce the following dual problem: seek $\psi \in H_0^1(\Omega)$ such that

$$(K\nabla v, \nabla\psi)_{0,\Omega} + (\beta\cdot\nabla v, \psi)_{0,\Omega} + (\mu v, \psi)_{0,\Omega} = (v, u - u_h)_{0,\Omega} \quad \forall v \in H_0^1(\Omega). \tag{2.29}$$

We assume that elliptic regularity holds in the broken H^2-norm, namely that

$$\|\psi\|_{H^2(\mathcal{T}_h)} \lesssim \lambda_{m,K}^{-1}\|u-u_h\|_{0,\Omega}. \tag{2.30}$$

When K is uniform, it is well-known that the convexity of Ω is sufficient to guarantee (2.30). This is no longer the case if K is discontinuous. In this case, (2.30) implicitly amounts to additional assumptions on the distribution of K inside Ω.

Theorem 2.5. *In the above framework,*

$$\|u-u_h\|_{0,\Omega} \le \frac{\lambda_{M,K}^{\frac12}}{\lambda_{m,K}} h \left(\|u-u_h\|_{h,B} + \inf_{w_h\in V_h} \|u-w_h\|_{h,B_+} \right) \tag{2.31}$$

where for all $v \in V(h)$,

$$\|v\|_{h,B_+} = \|v\|_{h,B} + \left(\sum_{T\in\mathcal{T}_h} h_T^2 \|\nabla_h v\|_{0,T}^2 \right)^{\frac12} + \left(\sum_{T\in\mathcal{T}_h} h_T \|\kappa\nabla_h v\|_{0,\partial T}^2 \right)^{\frac12}. \tag{2.32}$$

Proof. Step (i): observe that for all $v \in V(h)$, using (2.8) yields

$$\begin{aligned} B_h(v,\psi) &= (K\nabla_h v, \nabla\psi)_{0,\Omega} + ((\mu-\nabla\cdot\beta)v,\psi)_{0,\Omega} - (v,\beta\cdot\nabla\psi)_{0,\Omega} - \sum_{F\in\mathcal{F}_h} (n_F^t\{K\nabla\psi\}_\omega, [\![v]\!])_{0,F} \\ &= \sum_{T\in\mathcal{T}_h} (v, -\nabla\cdot(K\nabla\psi) - \beta\cdot\nabla\psi + (\mu-\nabla\cdot\beta)\psi)_{0,T} = (v,u-u_h)_{0,\Omega}. \end{aligned} \tag{2.33}$$

Step (ii): define on $V(h)$ the norm

$$\|v\|_{h,1} = \|v\|_{h,\frac12} + \left(\sum_{T\in\mathcal{T}_h} h_T^{-2}\|v\|_{0,T}^2 \right)^{\frac12} \tag{2.34}$$

and let us prove that for all $(v,w) \in V(h)\times V(h)$,

$$|B_h(v,w)| \lesssim \|v\|_{h,B_+}\|w\|_{h,1}. \tag{2.35}$$

Indeed, indicating by T_i, $1\le i\le 7$, the seven terms on the right-hand side of (2.9), and proceeding as in the proof of Lemma 2.3, it is clear that $\sum_{i\neq 3}|T_i| \lesssim \|v\|_{h,B_+}\|w\|_{h,\frac12}$. Moreover,

$$|T_3| = |(\beta\cdot\nabla_h v, w)_{0,\Omega}| \lesssim \sum_{T\in\mathcal{T}_h} \|\nabla_h v\|_{0,T}\|w\|_{0,T} = \sum_{T\in\mathcal{T}_h} h_T\|\nabla_h v\|_{0,T} h_T^{-1}\|w\|_{0,T} \le \|v\|_{h,B_+}\|w\|_{h,1}.$$

Hence, (2.35) holds.
Step (iii): taking $v = u - u_h$ in (2.33), applying Lemma 2.2 and using (2.35) yields for all $\psi_h \in V_h$,

$$\|u - u_h\|_{0,\Omega}^2 = B_h(u - u_h, \psi) = B_h(u - u_h, \psi - \psi_h) \lesssim \|u - u_h\|_{h,B_+} \|\psi - \psi_h\|_{h,1}.$$

Using standard interpolation results leads to

$$\inf_{\psi_h \in V_h} \|\psi - \psi_h\|_{h,1} \lesssim \lambda_{M,K}^{\frac{1}{2}} h \|\psi\|_{H^2(\mathcal{T}_h)},$$

and taking into account (2.30) yields

$$\|u - u_h\|_{0,\Omega} \lesssim \frac{\lambda_{M,K}^{\frac{1}{2}}}{\lambda_{m,K}} h \|u - u_h\|_{h,B_+}. \tag{2.36}$$

Using the inverse inequalities (2.17) and (2.18), we infer that for all $v_h \in V_h$,

$$\|v_h\|_{h,B_+} \lesssim \|v_h\|_{h,B} + \|v_h\|_{0,\Omega} + \|\kappa \nabla_h v_h\|_{0,\Omega} \lesssim \|v_h\|_{h,B}. \tag{2.37}$$

Applying the triangle inequality together with (2.37) leads to

$$\begin{aligned} \|u - u_h\|_{h,B_+} &\leq \|u - w_h\|_{h,B_+} + \|u_h - w_h\|_{h,B_+} \\ &\lesssim \|u - w_h\|_{h,B_+} + \|u_h - w_h\|_{h,B} \\ &\lesssim \|u - w_h\|_{h,B_+} + \|u - u_h\|_{h,B}, \end{aligned} \tag{2.38}$$

where w_h is arbitrary in V_h. Substituting (2.38) into (2.36) yields (2.31). □

Corollary 2.6. *If the exact solution u is in $H^{p+1}(\mathcal{T}_h)$, then*

$$\|u - u_h\|_{0,\Omega} \lesssim \frac{\lambda_{M,K}}{\lambda_{m,K}} h^{p+1} \|u\|_{H^{p+1}(\mathcal{T}_h)}. \tag{2.39}$$

Proof. Use Theorem 2.5 and standard approximation properties of V_h. □

2.4 Error analysis for the advective derivative

When the diffusivity takes small values, it is no longer possible to control the advective derivative by means of Theorem 2.1. The goal of this section is to obtain a control of

the error in the advective derivative that is possibly robust with respect to the diffusivity. Define on $V(h)$ the norm

$$\|v\|_{h,B\beta} = \|v\|_{h,B} + \|v\|_{h,\beta}, \tag{2.40}$$

where

$$\|v\|_{h,\beta} = \left(\sum_{T\in\mathcal{T}_h} h_T \|\beta\cdot\nabla_h v\|_{0,T}^2 \right)^{\frac{1}{2}}. \tag{2.41}$$

To prove a convergence result in the $\|\cdot\|_{h,B\beta}$-norm, the first step is to derive a stability property for the SWIP bilinear form B_h in this norm.

Lemma 2.7. (Stability) *Define*

$$\forall T\in\mathcal{T}_h, \qquad \Delta_{K,T} = \begin{cases} 1 & \text{if } \|\beta\|_{[L^\infty(T)]^d} \gtrsim \frac{\lambda_{M,T}}{h_T}, \\ \frac{\lambda_{M,T}}{\lambda_{m,T}} & \text{otherwise}, \end{cases} \tag{2.42}$$

where $\lambda_{M,T}$ and $\lambda_{m,T}$ are respectively the maximum and the minimum eigenvalue of $K|_T$. Set $\Delta_K = \max_{T\in\mathcal{T}_h} \Delta_{K,T}$. Then,

$$\inf_{v_h\in V_h\setminus\{0\}} \sup_{w_h\in V_h\setminus\{0\}} \frac{B_h(v_h,w_h)}{\|v_h\|_{h,B\beta}\|w_h\|_{h,B\beta}} \gtrsim \Delta_K^{-1}. \tag{2.43}$$

Remark. We stress the fact that the inf-sup condition is robust in the isotropic case and in the anisotropic case if the cell Péclet numbers evaluated with the largest eigenvalue of the diffusivity tensor are large enough. Note also that the anisotropies are local to the mesh element, i.e., ratios of eigenvalues between adjacent elements are not considered. To achieve this result, the key point (see the control of $|[\![\pi_h]\!]|_{\gamma_{K,F}}^2$ in the proof below) is that the choice (2.15) for the weights yields $\gamma_{K,F} \le \inf(\delta_{K,F-}, \delta_{K,F+})$.

Proof. Let $v_h \in V_h$ and set $\mathbf{S} = \sup_{w_h\in V_h\setminus\{0\}} \frac{B_h(v_h,w_h)}{\|w_h\|_{h,B\beta}}$. We want to prove that $\|v_h\|_{h,B\beta} \lesssim \Delta_K\mathbf{S}$.
Step (i): owing to Lemma 2.1, we infer that

$$\|v_h\|_{h,B}^2 \lesssim \mathbf{S}\|v_h\|_{h,B\beta}, \tag{2.44}$$

so it only remains to control the advective derivative in $\|v_h\|_{h,B\beta}$.

Step (ii): let $\pi_h \in V_h$ be such that for all $T \in \mathcal{T}_h$ $\pi_h|_T = h_T\overline{\beta}\cdot\nabla_h v_h$ where $\overline{\beta}$ is defined in the proof of Lemma 2.3. Let us prove that

$$\|\pi_h\|_{h,B\beta} \lesssim \Delta_K^{\frac{1}{2}}\|v_h\|_{h,B\beta}. \tag{2.45}$$

The inverse inequality (2.18) and the regularity of β yield for all $T \in \mathcal{T}_h$,

$$\|\pi_h\|_{0,T} \lesssim h_T\|\beta\cdot\nabla_h v_h\|_{0,T} + h_T\|v_h\|_{0,T}, \tag{2.46}$$

while the inverse inequality (2.17) yields for all $F \in \mathcal{F}_h$

$$|[\![\pi_h]\!]|^2_{\gamma_{\beta,F}} \lesssim \sum_{T\in\mathcal{T}(F)} \|\pi_h\|^2_{0,\partial T} \lesssim \sum_{T\in\mathcal{T}(F)} \left(h_T\|\beta\cdot\nabla_h v_h\|^2_{0,T} + h_T\|v_h\|^2_{0,T}\right).$$

Hence, since $\Delta_K \geq 1$,

$$\|\pi_h\|_{0,\Omega} + |[\![\pi_h]\!]|_{\gamma_{\beta,F}} \lesssim \|v_h\|_{h,B\beta} \leq \Delta_K^{\frac{1}{2}}\|v_h\|_{h,B\beta}.$$

Let us estimate $h_F^{-\frac{1}{2}}|[\![\pi_h]\!]|_{\gamma_{K,F}}$ for all $F \in \mathcal{F}_h$. Observe first that $\gamma_{K,F} = \omega_{T^{\mp}(F),F}\delta_{K,F^{\mp}} \leq \delta_{K,F^{\mp}}$ if $F \in \mathcal{F}_h^i$ and $\gamma_{K,F} = \delta_{K,F}$ if $F \in \mathcal{F}_h^{\partial\Omega}$. Hence, if there is a $T \in \mathcal{T}_h(F)$ such that $\|\beta\|_{[L^\infty(T)]^d} \gtrsim \frac{\lambda_{M,T}}{h_T}$, then

$$h_F^{-1}|[\![\pi_h]\!]|^2_{\gamma_{K,F}} \leq h_F^{-1}\lambda_{M,T}\|[\![\pi_h]\!]\|^2_{0,F} \leq \sum_{T\in\mathcal{T}(F)} \left(h_T\|\beta\cdot\nabla_h v_h\|^2_{0,T} + h_T\|v_h\|^2_{0,T}\right).$$

Otherwise, for all $F \in \mathcal{F}_h^i$,

$$\begin{aligned} h_F^{-1}\gamma_{K,F}[\![\pi_h]\!]^2 &\lesssim h_F\gamma_{K,F}\left(((\overline{\beta}\cdot\nabla_h v_h)^-)^2 + ((\overline{\beta}\cdot\nabla_h v_h)^+)^2\right) \\ &\lesssim h_F\left(\delta_{K,F^-}((\overline{\beta}\cdot\nabla_h v_h)^-)^2 + \delta_{K,F^+}((\overline{\beta}\cdot\nabla_h v_h)^+)^2\right), \end{aligned}$$

and similarly for $F \in \mathcal{F}_h^{\partial\Omega}$. Hence, using the trace inverse inequality (2.17),

$$h_F^{-1}|[\![\pi_h]\!]|^2_{\gamma_{K,F}} \lesssim \sum_{T\in\mathcal{T}(F)} \lambda_{M,T}\|\nabla_h v_h\|^2_{0,T} \lesssim \sum_{T\in\mathcal{T}(F)} \frac{\lambda_{M,T}}{\lambda_{m,T}}\|\kappa\nabla_h v_h\|^2_{0,T}.$$

Thus, $|[\![\pi_h]\!]|_{\gamma_F} \lesssim \Delta_K^{\frac{1}{2}}\|v_h\|_{h,B\beta}$. Furthermore, since κ is piecewise constant,

$$\|\kappa\nabla_h\pi_h\|_{0,T} = h_T\|\overline{\beta}\cdot\nabla_h(\kappa\nabla_h v_h)\|_{0,T} \lesssim \|\kappa\nabla_h v_h\|_{0,T},$$

implying that $\|\kappa\nabla_h\pi_h\|_{0,\Omega} \lesssim \|v_h\|_{h,B}$. Finally, the advective derivative of π_h is controlled by

$$\|\pi_h\|_{h,\beta}^2 \lesssim \sum_{T\in\mathcal{T}_h} h_T^{-1}\|\pi_h\|_{0,T}^2 \lesssim \|v_h\|_{h,B\beta}^2,$$

owing to (2.46). This proves (2.45).
Step (iii): we can now examine the term $\|v_h\|_{h,\beta}^2$ by making use of (2.9):

$$\begin{aligned}\|v_h\|_{h,\beta}^2 &= B_h(v_h,\pi_h) - (K\nabla_h v_h,\nabla_h\pi_h)_{0,\Omega} - (\mu v_h,\pi_h)_{0,\Omega}\\ &\quad+\sum_{T\in\mathcal{T}_h}(\beta\cdot\nabla_h v_h, h_T\beta\cdot\nabla_h v_h-\pi_h)_{0,T} + \sum_{F\in\mathcal{F}_h}(\beta\cdot n_F\{\pi_h\},[\![v_h]\!])_{0,F}\\ &\quad+\sum_{F\in\mathcal{F}_h}\big((n_F^t\{K\nabla_h v_h\}_\omega,[\![\pi_h]\!])_{0,F}+(n_F^t\{K\nabla_h\pi_h\}_\omega,[\![v_h]\!])_{0,F}-(\gamma_F[\![v_h]\!],[\![\pi_h]\!])_{0,F}\big)\\ &= B_h(v_h,\pi_h)+T_1+T_2+T_3+T_4+T_5+T_6+T_7.\end{aligned}$$

We observe that

$$|B_h(v_h,\pi_h)| \le \mathbf{S}\|\pi_h\|_{h,B\beta} \le \mathbf{S}\Delta_K^{\frac12}\|v_h\|_{h,B\beta}.$$

It is also clear that, using (2.44),

$$|T_1|+|T_2|+|T_5|+|T_6|+|T_7| \lesssim \|v_h\|_{h,B}\|\pi_h\|_{h,B} \lesssim \mathbf{S}^{\frac12}\Delta_K^{\frac12}\|v_h\|_{h,B\beta}^{\frac32}.$$

Furthermore, using the inverse inequality (2.17) together with (2.46) yields

$$\begin{aligned}|T_4| \lesssim |[\![v_h]\!]|_{\gamma_{\beta,F}}\left(\sum_{T\in\mathcal{T}_h}\|\pi_h\|_{0,\partial T}^2\right)^{\frac12} &\lesssim |[\![v_h]\!]|_{\gamma_{\beta,F}}\left(\sum_{T\in\mathcal{T}_h}h_T^{-1}\|\pi_h\|_{0,T}^2\right)^{\frac12}\\ &\lesssim \|v_h\|_{h,B}\|v_h\|_{h,B\beta} \lesssim \mathbf{S}^{\frac12}\|v_h\|_{h,B\beta}^{\frac32}.\end{aligned}$$

Finally,

$$\begin{aligned}|T_3| &\le \sum_{T\in\mathcal{T}_h} h_T|(\beta\cdot\nabla_h v_h,(\beta-\overline{\beta})\cdot\nabla_h v_h)_{0,T}| \lesssim \sum_{T\in\mathcal{T}_h} h_T^2\|\beta\cdot\nabla_h v_h\|_{0,T}\|\nabla_h v_h\|_{0,T}\\ &\lesssim \sum_{T\in\mathcal{T}_h} h_T\|\beta\cdot\nabla_h v_h\|_{0,T}\|v_h\|_{0,T} \lesssim \|v_h\|_{h,B\beta}\|v_h\|_{0,\Omega} \lesssim \mathbf{S}^{\frac12}\|v_h\|_{h,B\beta}^{\frac32}.\end{aligned}$$

Hence,

$$\begin{aligned}\|v_h\|_{h,B\beta}^2 &\lesssim \|v_h\|_{h,B}^2 + \|v_h\|_{h,\beta}^2 \\ &\lesssim \mathbf{S}\|v_h\|_{h,B\beta} + \mathbf{S}\Delta_K^{\frac{1}{2}}\|v_h\|_{h,B\beta} + \mathbf{S}^{\frac{1}{2}}\Delta_K^{\frac{1}{2}}\|v_h\|_{h,B\beta}^{\frac{3}{2}} + \mathbf{S}^{\frac{1}{2}}\|v_h\|_{h,B\beta}^{\frac{3}{2}} \\ &\lesssim \mathbf{S}\Delta_K^{\frac{1}{2}}\|v_h\|_{h,B\beta} + \mathbf{S}^{\frac{1}{2}}\Delta_K^{\frac{1}{2}}\|v_h\|_{h,B\beta}^{\frac{3}{2}},\end{aligned}$$

where we have used the fact that $\Delta_K \geq 1$ in the last step. Applying twice Young's inequality yields the desired result. □

Proceeding as above, the following result is readily inferred:

Theorem 2.8. *In the above framework,*

$$\|u - u_h\|_{h,B\beta} \lesssim \Delta_K \inf_{v_h \in V_h} \|u - v_h\|_{h,\frac{1}{2}\beta}, \tag{2.47}$$

where, for all $v \in V(h)$,

$$\|v\|_{h,\frac{1}{2}\beta} = \|v\|_{h,B\beta} + \left(\sum_{T\in\mathcal{T}_h} \|v\|_{0,\partial T}^2\right)^{\frac{1}{2}} + \left(\sum_{T\in\mathcal{T}_h} h_T\|\kappa\nabla_h v\|_{0,\partial T}^2\right)^{\frac{1}{2}}. \tag{2.48}$$

Remark. Estimate (2.47) yields an error upper bound on the advective derivative with a constant depending on Δ_K. Robustness is recovered whenever $\Delta_K = 1$, i.e., when working with an isotropic diffusivity tensor or when the cell Péclet numbers evaluated with the largest eigenvalue of the diffusivity tensor are large enough. Furthermore, if $u \in H^{p+1}(\mathcal{T}_h)$, the upper bound converges as $h^{p+\frac{1}{2}}$, which is optimal.

2.5 Numerical tests

2.5.1 A test case with discontinuous coefficients

To verify the convergence of the SWIP method and to make quantitative comparisons between this and other IP methods, we consider the test problem proposed in [25], featuring discontinuous coefficients and where the exact solution is known analytically. We split the domain $\Omega = [0,1]\times[0,1]$ into two subdomains: $\Omega_1 = [0,\frac{1}{2}]\times[0,1]$, $\Omega_2 = [\frac{1}{2},1]\times[0,1]$. The diffusivity tensor K is constant within each subdomain, and defined as

$$K(x,y) = \begin{pmatrix} \epsilon(x) & 0 \\ 0 & 1.0 \end{pmatrix}$$

where $\epsilon(x)$ is a discontinuous function across the interface $x = \frac{1}{2}$. Indicating with the subscript 1 (resp. 2) the restriction to the subdomain Ω_1 (resp. Ω_2), we will consider different values of ϵ_1, while ϵ_2 is set equal to 1. Letting $\beta = (1,0)^t$, $\mu = 0$ and $f = 0$, the exact solution is independent of the y-coordinate, and is exponential with respect to the x-coordinate. The following conditions must be satisfied at the interface between the two subdomains:

$$\lim_{x\to\frac{1}{2}^-} u(x,y) = \lim_{x\to\frac{1}{2}^+} u(x,y), \text{ and } \lim_{x\to\frac{1}{2}^-} -\epsilon_1\partial_x u(x,y) = \lim_{x\to\frac{1}{2}^+} -\partial_x u(x,y).$$

Setting $u(0,y) = 1$, $u(1,y) = 0$ and applying the matching conditions, we obtain the value of the exact solution at the interface:

$$u\left(\tfrac{1}{2},y\right) = \frac{\exp(\frac{1}{2\epsilon_1})}{1-\exp(\frac{1}{2\epsilon_1})}\left(\frac{\exp(\frac{1}{2\epsilon_1})}{1-\exp(\frac{1}{2\epsilon_1})} + \frac{1}{1-\exp(\frac{1}{2})}\right)^{-1}.$$

As a result, the exact solution in each subdomain can be expressed as

$$u_1(x,y) = \frac{u(\frac{1}{2},y) - \exp(\frac{1}{2\epsilon_1}) + (1-u(\frac{1}{2},y))\exp(\frac{x}{\epsilon_1})}{1-\exp(\frac{1}{2\epsilon_1})},$$
$$u_2(x,y) = \frac{-\exp(\frac{1}{2})u(\frac{1}{2},y) + u(\frac{1}{2},y)\exp(x-\frac{1}{2})}{1-\exp(\frac{1}{2})}.$$

h	$\|u-u_h\|_{h,B}$	$\|u-u_h\|_{h,\beta}$	$\|u-u_h\|_{0,\Omega}$
0.1000	1.62e-01	1.49e-01	6.94e-03
0.0500	7.96e-02	5.45e-02	2.11e-03
0.0250	3.67e-02	1.87e-02	4.80e-04
0.0125	1.70e-02	6.37e-03	1.21e-04
order	1.11	1.55	1.98

Table 2.1: Convergence rates of the SWIP method, $p = 1$

h	$\|u-u_h\|_{h,B}$	$\|u-u_h\|_{h,\beta}$	$\|u-u_h\|_{0,\Omega}$
0.1000	2.31e-02	2.15e-02	6.80e-04
0.0500	4.63e-03	3.31e-03	4.29e-05
0.0250	1.17e-03	5.93e-04	5.20e-06
0.0125	2.95e-04	1.05e-04	6.41e-07
order	1.99	2.49	3.02

Table 2.2: Convergence rates of the SWIP method, $p = 2$

To assess the accuracy of the SWIP method with respect to the mesh-size, we consider a family of uniform triangulations $\{\mathcal{T}_h\}_{h>0}$ which are conforming with respect to the interface between Ω_1 and Ω_2. These triangulations are obtained starting from a uniform partition of $\partial\Omega$ in sub-intervals of length $h = 0.1$, $h = 0.05$, $h = 0.025$ and $h = 0.0125$ respectively. The value of the penalty parameter α is henceforth set to $\alpha = 1.0$ for $\mathbb{P}_1$ elements and $\alpha = 4.0$ for $\mathbb{P}_2$ elements. The numerical results obtained with $\epsilon_1 = 0.1$ are reported in Tables 2.1 and 2.2, where the order of convergence is computed with respect to the last two rows of each table. We observe that the SWIP method exhibits the orders of convergence predicted by the theory.

method	$\|u-u_h\|_{h,B}$	$\|u-u_h\|_{h,\beta}$	$\|u-u_h\|_{0,\Omega}$	M
SWIP	1.583e-01	1.505e-01	4.586e-03	9.555e-04
IP-A	1.483e-01	1.403e-01	5.153e-03	5.882e-03
IP-B	1.338e-01	1.378e-01	5.903e-03	5.882e-03

Table 2.3: Comparison of SWIP and IP methods: $\epsilon_1 = 5\text{e-}2$, $p = 1$

method	$\|u-u_h\|_{h,B}$	$\|u-u_h\|_{h,\beta}$	$\|u-u_h\|_{0,\Omega}$	M
SWIP	4.917e-01	1.280	1.474e-02	6.594e-02
IP-A	5.886e-01	1.303	4.973e-02	4.373e-01
IP-B	6.625e-01	1.634	7.553e-02	4.173e-01

Table 2.4: Comparison of SWIP and IP methods: $\epsilon_1 = 5\text{e-}3$, $p = 1$

method	$\|u-u_h\|_{h,B}$	$\|u-u_h\|_{h,\beta}$	$\|u-u_h\|_{0,\Omega}$	M
SWIP	4.33e-01	1.44e+00	1.69e-02	6.72e-02
IP-A	6.05e-01	1.54e+00	3.77e-02	1.85e-01
IP-B	6.52e-01	1.71e+00	4.52e-02	1.86e-01

Table 2.5: Comparison of SWIP and IP methods: $\epsilon_1 = 5\text{e-}3$, $p = 2$

We have also compared the SWIP method with two IP methods. The first method (IP-A) corresponds to the SWIP method with weights $\omega_{T^{\mp}(F),F} = \frac{1}{2}$. The penalty parameter $\gamma_{K,F}$ is thus the arithmetic average of the diffusivity in the direction normal to the face. This method was analyzed in [53]. The second method (IP-B), proposed in [59], differs from IP-A in the choice of the penalty parameter: $\gamma_{K,F}$ is the arithmetic average of the maximum eigenvalue of K on the triangles sharing the face F. We consider a uniform triangulation $\mathcal{T}_h$ characterized by $h = 0.05$. The quantitative analysis is based on the norms $\|\cdot\|_{h,B}$, $\|\cdot\|_{h,\beta}$, $\|\cdot\|_{0,\Omega}$ and the indicator

$$M = \max(|\max_{\Omega}(u_h) - \max_{\Omega}(u)|, |\min_{\Omega}(u_h) - \min_{\Omega}(u)|) \tag{2.49}$$

which quantifies overshoots and undershoots of the calculated solution. The numerical results for $p = 1$ are found in Tables 2.3, 2.4, and in Figure 2.1. Table 2.3 deals with the case $\epsilon_1 = 5\text{e-}2$; the inner layer is not very sharp and is resolved by the meshes under consideration. The three methods deliver similar results for all the quantities of interest. As the inner layer becomes sharper ($\epsilon_1 = 5\text{e-}3$, Table 2.4), the SWIP scheme performs better than the other IP methods, especially in the L^2-norm and in the indicator M. The reason is that the weights permit sharper discontinuities in the calculated solution, leading to smaller oscillations in the internal layer, whereas the other IP methods force the discrete solution to be almost continuous. As can be observed in Figure 2.1, this limitation promotes instabilities in the neighborhood of the internal layer. The spurious oscillations generated in the case $\epsilon_1 = 5\text{e-}3$ lead to an overshoot of about 40%. The robustness of the SWIP method with respect to standard IP schemes is also confirmed by further numerical tests concerning vanishing values of ϵ_1 (Figure 2.2). Finally, Table 2.5 presents the results for $\epsilon_1 = 5\text{e-}3$ and $p = 2$. Here we use a coarser mesh yielding approximately the same number of degrees of freedom as in the tests with linear polynomials. Then, the same conclusion as for $p = 1$ can be reached. As the mesh is further refined (or the polynomial degree is further increased), the approximation space eventually becomes rich enough to completely capture the internal layer, and the three methods exhibit a similar behavior.

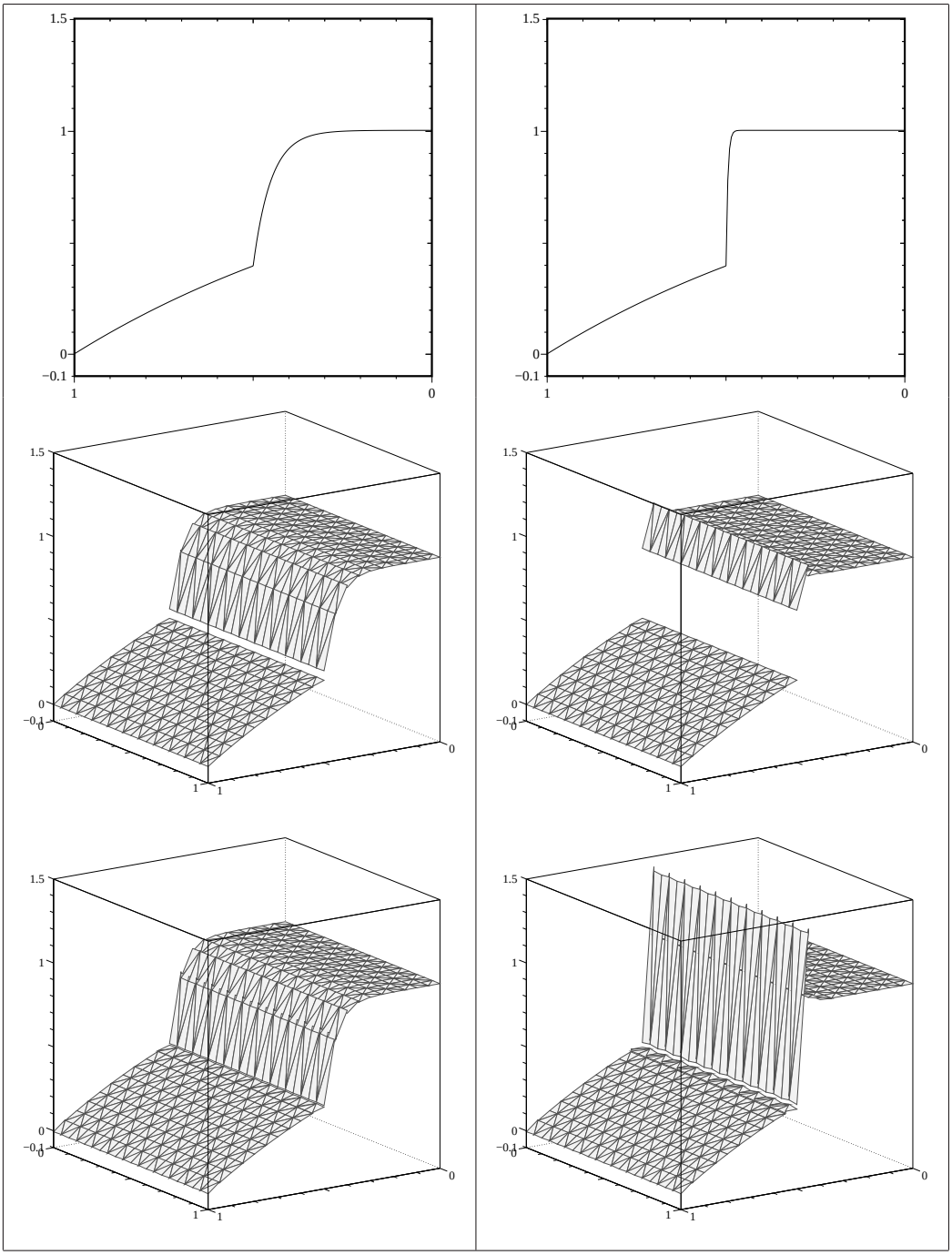


Figure 2.1: Graphical comparison between the methods SWIP and IP-A. The test case with $\epsilon_1 = 5\text{e-}2$ is reported on the left while the case with $\epsilon_1 = 5\text{e-}3$ is on the right. In both cases $\epsilon_2 = 1$. Each column shows the one-dimensional exact solution $u(x)$ of the test problem (top) and the numerical approximation u_h obtained with the methods SWIP (center) and IP-A (bottom), by means of piecewise-linear elements ($p = 1$). The case IP-B has been omitted since it is qualitatively equivalent to IP-A.

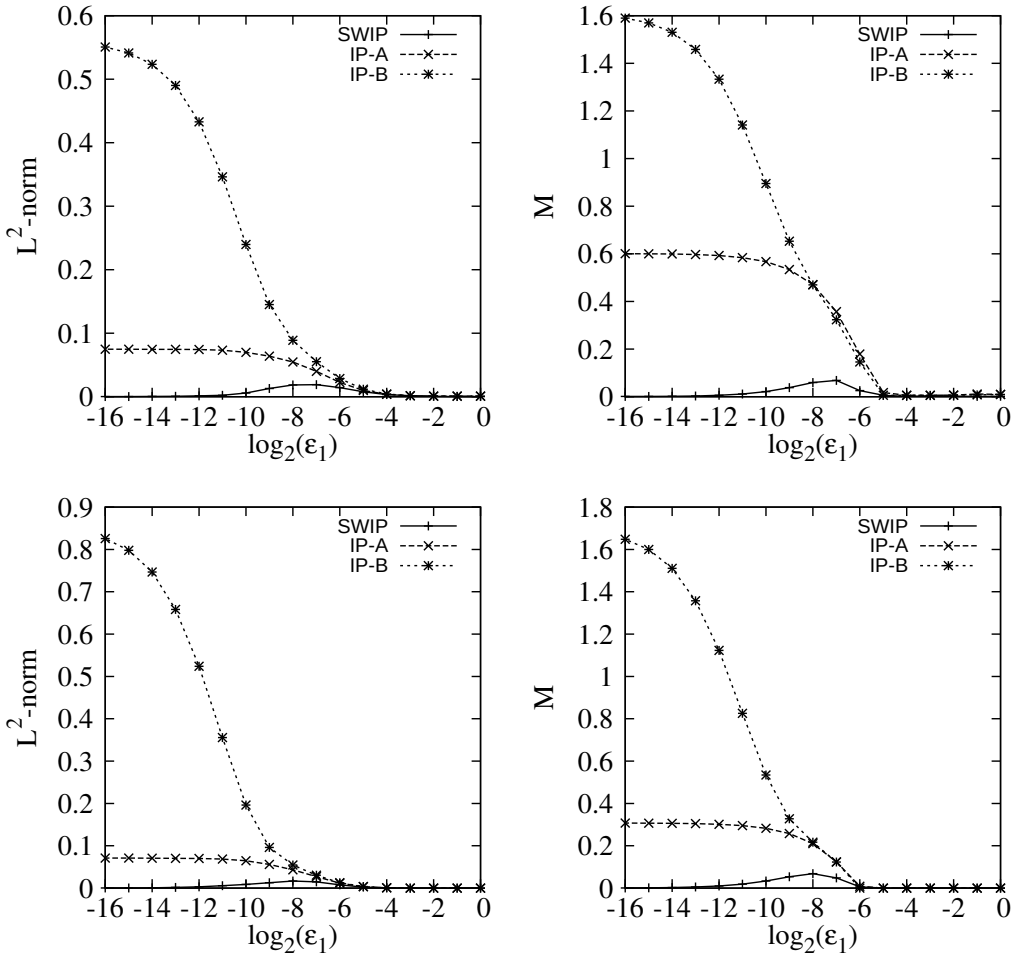


Figure 2.2: The norm $\|\cdot\|_{0,\Omega}$ and the indicator (2.49) (denoted by M) are plotted for the values $\epsilon_1 = 2^{-i},\ i = 0, \ldots, 16$. The methods SWIP, IP-A and IP-B are compared with respect to these indicators for linear (top) and quadratic elements (bottom).

2.5.2 A test case with genuine anisotropic properties

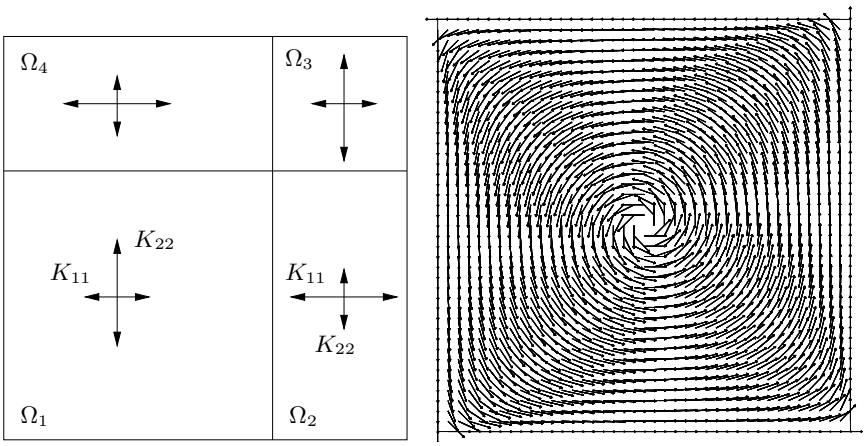


Figure 2.3: Test case with genuine anisotropic properties. On the left, an illustration of the domain and its subregions together with a synoptic description of the diffusivity tensor. The advection field β is shown on the right.

To conclude the sequence of numerical tests, we consider a test case with genuine anisotropic properties. Because of the complexity of the problem, it is not possible to compute analytically the exact solution. Consequently, the comparison between the SWIP and the IP methods will only be qualitative.

We consider the unit square $\Omega = [0,1] \times [0,1]$ split into four subdomains: $\Omega_1 = [0, \frac{2}{3}] \times [0, \frac{2}{3}]$, $\Omega_2 = [\frac{2}{3}, 1] \times [0, \frac{2}{3}]$, $\Omega_3 = [\frac{2}{3}, 1] \times [\frac{2}{3}, 1]$ and $\Omega_4 = [0, \frac{2}{3}] \times [\frac{2}{3}, 1]$. The diffusivity tensor K takes different values in each subregion:

$$K(x,y) = \begin{pmatrix} 1\text{e}{-6} & 0 \\ 0 & 1.0 \end{pmatrix} \text{ for } (x,y) \in \Omega_1,\ \Omega_3,$$

$$K(x,y) = \begin{pmatrix} 1.0 & 0 \\ 0 & 1\text{e}{-6} \end{pmatrix} \text{ for } (x,y) \in \Omega_2,\ \Omega_4.$$

The advection field is solenoidal and given by $\beta = (\beta_x, \beta_y)^t$ with $\beta_x = 40x(2y-1)(x-1)$ and $\beta_y = -40y(2x-1)(y-1)$. Unlike the previous test case, we note that the field β is neither constant nor orthogonal to the interfaces of discontinuity of K, but it is still oriented along the direction of increasing diffusivity, thus triggering internal layers. The forcing term only depends on the radial coordinate originating at the center of Ω in the form

$f(x,y) = 10^{-2}\exp(-(r-0.35)^2/0.005)$ with $r^2 = (x-0.5)^2 + (y-0.5)^2$; this corresponds to a Gaussian hill with center at $r = 0.35$. Finally, we choose $\mu = 1$. For the simulations, we consider a quasi-uniform mesh with $h = 0.025$. The mesh is conforming with respect to the discontinuities of K. A qualitative representation of the data is found in Figure 2.3.

In the left column of Figure 2.4 we compare the solutions obtained with the SWIP and the IP methods. The contour plots of the numerical solutions confirm that the methods at hand behave differently in the neighborhood of the interfaces where the tensor K is discontinuous. We observe that the SWIP scheme approximates the internal layers by means of jumps, while the IP schemes attempt to recover a numerical solution which is almost continuous. Since the computational mesh is insufficiently refined, the scheme IP-A generates some slight undershoots near the interfaces where K is discontinuous. For the IP-B method the oscillations generated by the approximation of the internal layer are much more evident and propagate quite far away from the interfaces. This behavior can be explained by observing that this type of penalty does not distinguish between the principal directions of the diffusivity tensor. Consequently, an excessive penalty is applied along the direction of low diffusivity.

To strengthen these conclusions, we also consider a numerical test where the advection field is the opposite of the one reported in Figure 2.3, i.e. it rotates clockwise. Following this advection field along the interfaces between subdomains, the diffusivity decreases. These conditions lead to an exact solution which is smooth in the neighborhood of the interfaces. In this case, the three methods are expected to behave similarly, as is confirmed by the numerical results reported in the right column of Figure 2.4.

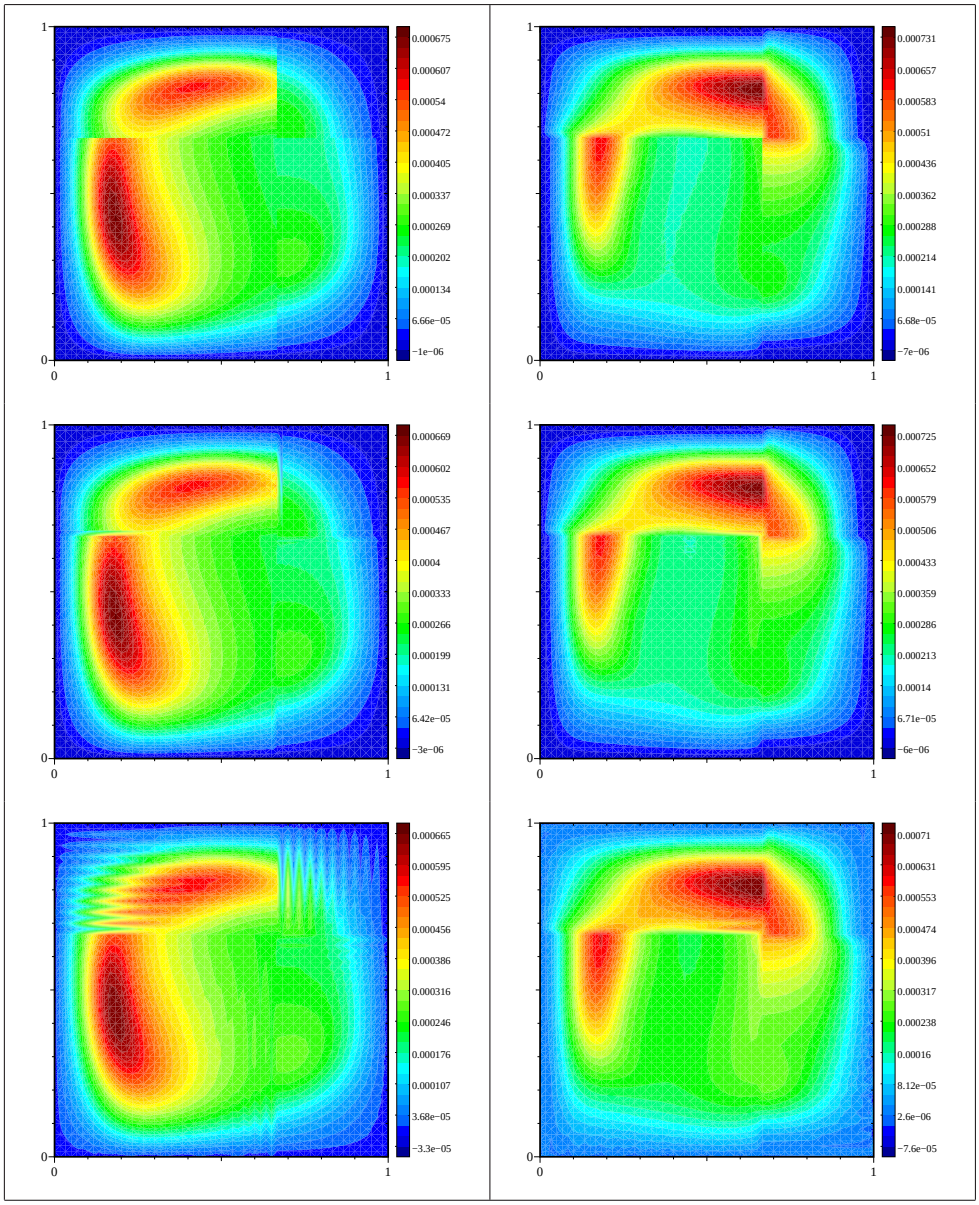


Figure 2.4: Test case with genuine anisotropic properties. The advection field rotates counterclockwise on the left (see figure 2.3) and clockwise on the right. The solution obtained by the SWIP scheme is shown on the top while those relative to the IP-A and IP-B methods are depicted below.

2.6 Concluding remarks

The SWIP method analyzed in this paper is a DG method with weighted averages designed to approximate satisfactorily advection-diffusion-reaction equations with anisotropic and locally small diffusivity. A thorough a priori error analysis has been carried out, yielding robust and optimal error estimates that have been supported by numerical evidence. The SWIP method is an interesting alternative to other IP methods since it can approximate more sharply under-resolved internal layers caused by locally small diffusivity.

Chapitre 3

A posteriori energy-norm error estimate

Submitted to Journal of Computational Mathematics under the title 'A posteriori energy-norm error estimates for advection-diffusion equations approximated by weighted interior penalty methods'.

Alexandre Ern[1] and Annette F. Stephansen[1,2]

Abstract: We propose and analyze a posteriori energy-norm error estimates for weighted interior penalty discontinuous Galerkin approximations to advection-diffusion-reaction equations with heterogeneous and anisotropic diffusion. The weights, which play a key role in the analysis, depend on the diffusion tensor and are used to formulate the consistency terms in the discontinuous Galerkin method. The error upper bounds, in which all the constants are specified, consist of three terms: a residual estimator which depends only on the elementwise fluctuation of the discrete solution residual, a diffusive flux estimator where the weights used in the method enter explicitly, and a non-conforming estimator which is nonzero because of the use of discontinuous finite element spaces. The three estimators can be bounded locally by the approximation error. A particular attention is given to the dependency on problem parameters of the constants in the local lower error bounds. For moderate advection, it is shown that full robustness with respect to diffusion heterogeneities is achieved owing to the specific design of the weights in the discontinuous Galerkin method, while diffusion anisotropies remain purely local and impact the constants through the square root of the condition number of the diffusion tensor. For dominant advection, it is shown,

[1]Cermics, Ecole des Ponts, ParisTech, 6 et 8 avenue Blaise Pascal, Champs sur Marne, 77455 Marne la Vallée Cedex 2, France.

[2]Andra, Parc de la Croix-Blanche, 1-7 rue Jean Monnet, 92298 Châtenay-Malabry cedex, France.

in the spirit of previous work by Verfürth on continuous finite elements, that the constants are bounded by the square root of the local Péclet number.

3.1 Introduction

In this work, we are interested in a posteriori energy-norm error estimates for a particular class of discontinuous Galerkin (DG) approximations of the advection-diffusion-reaction equation

$$\begin{cases} -\nabla\cdot(K\nabla u) + \beta\cdot\nabla u + \mu u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{3.1}$$

where for simplicity homogeneous Dirichlet boundary conditions are considered. Here, Ω is a polygonal domain in $\mathbb{R}^d$ with boundary $\partial\Omega$, $\mu \in L^\infty(\Omega)$, $\beta \in [L^\infty(\Omega)]^d$ with $\nabla\cdot\beta \in L^\infty(\Omega)$, $\tilde{\mu} := \mu - \frac{1}{2}\nabla\cdot\beta$ is assumed to be nonnegative, the diffusion tensor K is a symmetric, uniformly positive definite field in $[L^\infty(\Omega)]^{d,d}$ and $f \in L^2(\Omega)$. Owing to the above assumptions, (3.1) is well-posed.

DG methods received extensive interest in the past decade, in particular because of the flexibility they offer in the construction of approximation spaces using non-matching meshes and variable polynomial degrees. For diffusion problems, various DG methods have been analyzed, including the Symmetric Interior Penalty method [10, 15], the Nonsymmetric method with [90] or without [78] penalty, and the Local Discontinuous Galerkin method [33]; see [9] for a unified analysis. For linear hyperbolic problems (e.g., advection–reaction), one of the most common approaches is to use upwind fluxes to formulate the DG method [61, 68]. A unified theory of DG approximations encompassing elliptic and hyperbolic PDE's can be found in [43, 44]. The approximation of the advection-diffusion-reaction problem (3.1) using DG methods has been analyzed in [59] and more recently in [45] with a focus on the high Péclet regime with isotropic and uniform diffusion. The case of high contrasts in the diffusivity poses additional difficulties. Recently, a (Symmetric) Weighted Interior Penalty method has been proposed and analyzed to approximate satisfactorily (3.1) in this situation [51]. The key idea is to use weighted averages (depending on the normal diffusivities at the two mesh elements sharing a given interface) to formulate the consistency terms and to penalize the jumps of the discrete solution by a factor proportional to the harmonic mean of the neighboring normal diffusivities; the idea of using weighted interior penalties in this context can be traced back to [25].

The present paper addresses the a posteriori error analysis of the weighted interior penalty method. Many significant advances in the a posteriori error analysis of DG methods have been accomplished in the past few years. For energy-norm estimates, we refer to the pioneering work of Becker, Hansbo and Larson [20] and that of Karakashian and Pascal [62], while further developments can be found in the work of Ainsworth [3,4] regarding robustness with respect to diffusivity and that of Houston, Schötzau and Wihler [58] regarding the *hp*-analysis; see also [26,93]. Furthermore, for L^2-norm estimates, we mention the work of Becker, Hansbo and Stenberg [21], that of Rivière and Wheeler [87], and that of Castillo [29]. Broadly speaking, two approaches can be undertaken to derive a posteriori energy-norm error estimates; in [3,20,26], a Helmholtz decomposition of the error is used, following a technique introduced in [27,35], while the analysis in [58,62] relies more directly on identifying a conforming part in the discrete solution. The analysis presented herein will be closer to the latter approach. We also mention recent work relying on the reconstruction of a diffusive flux; see [50,67].

This paper is organized as follows. §3.2 presents the discrete setting, including the weighted interior penalty bilinear form used to formulate the discrete problem. §3.3 contains the main results of this work. The starting point is the abstract framework for a posteriori error estimates presented in §3.3.1 and which is closely inspired by the work of Vohralík for mixed finite element discretizations [102]. Then, §3.3.2 addresses the case of pure diffusion with heterogeneous and possibly anisotropic diffusivity. We derive an upper bound for the error consisting of three error indicators, i.e. a residual, a diffusive flux and a non-conforming one. This form is similar to that obtained in previous work. The key point however is that the diffusive flux error indicators also provide local lower error bounds that are fully robust with respect to diffusivity heterogeneities and that depend on the local (elementwise) degree of anisotropy; see Propositions 3.4 and 3.5. A key ingredient to obtain this result is the use of weighted averages in writing the consistency term. §3.3.3 extends the previous analysis to the advection-diffusion-reaction problem. Here, the focus is set on achieving a certain degree of robustness in the high Péclet regime, namely that achieved by Verfürth [98] for a posteriori energy-norm error estimates with conforming finite elements and SUPG stabilization. Although these estimates are not independent of the Péclet number (see, e.g., [99] for fully robust estimates with suitable norm modification), their present extension to DG methods constitutes the first results of this type. Finally, numerical results are presented in §3.4.

3.2 The discrete setting

Let $\{\mathcal{T}_h\}_{h>0}$ be a shape-regular family of affine triangulations covering exactly the polygonal domain Ω. The meshes $\mathcal{T}_h$ may possess hanging nodes, as long as the number of hanging nodes per mesh element is uniformly bounded. For meshes with hanging nodes, the shape-regularity must hold for a hierarchical refinement of the mesh without hanging nodes. This assumption is needed in the local lower error bounds when using the approximation properties of the Oswald interpolate (see (3.35)–(3.36) below) and when working with edge bubble functions (see, e.g., the proof of Proposition 3.5). A generic element in $\mathcal{T}_h$ is denoted by T, h_T denotes the diameter of T and n_T its outward unit normal. Let an integer $p \geq 1$. We consider the usual DG approximation space

$$V_h = \{v_h \in L^2(\Omega); \forall T \in \mathcal{T}_h, v_h|_T \in \mathbb{P}_p\}, \tag{3.2}$$

where $\mathbb{P}_p$ is the set of polynomials of degree less than or equal to p. The L^2-scalar product and its associated norm on a region $R \subset \Omega$ are indicated by the subscript $0, R$. For $s \geq 1$, a norm (semi-norm) with the subscript s, R designates the usual norm (semi-norm) in $H^s(R)$. For $s \geq 1$, $H^s(\mathcal{T}_h)$ denotes the usual broken Sobolev space on $\mathcal{T}_h$. For $v \in H^1(\mathcal{T}_h)$, $\nabla_h v$ denotes the piecewise gradient of v, that is, $\nabla_h v \in [L^2(\Omega)]^d$ and for all $T \in \mathcal{T}_h$, $(\nabla_h v)|_T = \nabla(v|_T)$.

We say that F is an interior face of the mesh if there are $T^-(F)$ and $T^+(F)$ in $\mathcal{T}_h$ such that $F = T^-(F) \cap T^+(F)$. We set $\mathcal{T}(F) = \{T^-(F), T^+(F)\}$ and let n_F be the unit normal vector to F pointing from $T^-(F)$ towards $T^+(F)$. The analysis hereafter does not depend on the arbitrariness of this choice. Similarly, we say that F is a boundary face of the mesh if there is $T^-(F) \in \mathcal{T}_h$ such that $F = T^-(F) \cap \partial\Omega$. We set $\mathcal{T}(F) = \{T^-(F)\}$ and let n_F coincide with the outward normal to $\partial\Omega$. All the interior (resp., boundary) faces of the mesh are collected into the set $\mathcal{F}_h^i$ (resp., $\mathcal{F}_h^{\partial\Omega}$) and we let $\mathcal{F}_h = \mathcal{F}_h^i \cup \mathcal{F}_h^{\partial\Omega}$. For $T \in \mathcal{T}_h$, $\mathcal{F}_T$ denotes the set of its faces and $\widetilde{\mathcal{F}}_T$ the set of mesh faces that share at least a vertex with T. Henceforth, we shall often deal with functions that are double-valued on $\mathcal{F}_h^i$ and single-valued on $\mathcal{F}_h^{\partial\Omega}$. This is the case, for instance, of functions in V_h. On interior faces, when the two branches of the function in question, say v, are associated with restrictions to the neighboring elements $T^{\mp}(F)$, these branches are denoted by $v^{\mp}$ and the jump of v across F is defined as

$$[\![v]\!]_F = v^- - v^+. \tag{3.3}$$

We set $[\![v]\!]_F = v|_F$ on boundary faces. On an interior face $F \in \mathcal{F}_h^i$, we also define the standard (arithmetic) average as $\{v\}_F = \frac{1}{2}(v^- + v^+)$. The subscript F in the above jumps

and averages is omitted if there is no ambiguity. We define the weighted average of a two-valued function v on an interior face $F \in \mathcal{F}_h^i$ as

$$\{v\}_\omega = \omega_{T^-(F),F} v^- + \omega_{T^+(F),F} v^+, \tag{3.4}$$

where the weights are defined as

$$\omega_{T^-(F),F} = \frac{\delta_{K,F+}}{\delta_{K,F+} + \delta_{K,F-}}, \qquad \omega_{T^+(F),F} = \frac{\delta_{K,F-}}{\delta_{K,F+} + \delta_{K,F-}}, \tag{3.5}$$

with $\delta_{K,F\mp} = n_F(K|_{T^\mp})n_F$. We extend the above definitions to boundary faces by formally setting $\omega_{T^-(F),F} = 1$ and $\omega_{T^+(F),F} = 0$. For the standard average, it is instead more convenient to set $\{v\}_F = \frac{1}{2} v|_F$ on boundary faces.

The weak formulation of (3.1) consists of finding $u \in V := H_0^1(\Omega)$ such that

$$B(u, v) = (f, v)_{0,\Omega}, \qquad \forall v \in V, \tag{3.6}$$

with the bilinear form

$$B(v, w) = (K\nabla_h v, \nabla_h w)_{0,\Omega} + (\beta \cdot \nabla_h v, w)_{0,\Omega} + (\mu v, w)_{0,\Omega}. \tag{3.7}$$

Piecewise gradients are used so as to extend the domain of B to functions in $V + V_h$. The energy norm is

$$\|v\|_B^2 = \sum_{T \in \mathcal{T}_h} \|v\|_{B,T}^2, \qquad \|v\|_{B,T}^2 = (K\nabla_h v, \nabla_h v)_{0,T} + (\tilde{\mu} v, v)_{0,T}. \tag{3.8}$$

The discrete problem consists of finding $u_h \in V_h$ such that

$$B_h(u_h, v_h) = (f, v_h)_{0,\Omega}, \qquad \forall v_h \in V_h, \tag{3.9}$$

with the bilinear form

$$\begin{aligned} B_h(v, w) = {} & (K\nabla_h v, \nabla_h w)_{0,\Omega} + ((\mu - \nabla \cdot \beta) v, w)_{0,\Omega} - (v, \beta \cdot \nabla_h w)_{0,\Omega} \\ & + \sum_{F \in \mathcal{F}_h} [(\gamma_F [\![v]\!], [\![w]\!])_{0,F} - (n_F^t \{K\nabla_h v\}_\omega, [\![w]\!])_{0,F} - \theta (n_F^t \{K\nabla_h w\}_\omega, [\![v]\!])_{0,F}] \\ & + \sum_{F \in \mathcal{F}_h} (\beta \cdot n_F \{v\}, [\![w]\!])_{0,F}. \end{aligned} \tag{3.10}$$

The penalty parameter γ_F is defined for all $F \in \mathcal{F}_h$ as $\gamma_F = \alpha h_F^{-1} \gamma_{K,F} + \gamma_{\beta,F}$ with

$$\forall F \in \mathcal{F}_h^i, \qquad \gamma_{K,F} = \frac{\delta_{K,F+}\delta_{K,F-}}{\delta_{K,F+} + \delta_{K,F-}}, \tag{3.11}$$

$$\forall F \in \mathcal{F}_h^{\partial\Omega}, \qquad \gamma_{K,F} = \delta_{K,F}, \tag{3.12}$$

$$\forall F \in \mathcal{F}_h, \qquad \gamma_{\beta,F} = \tfrac{1}{2} |\beta \cdot n_F|, \tag{3.13}$$

and α is a positive parameter (α can also vary from face to face). Finally, the parameter θ can take values in $\{-1, 0, +1\}$. The particular value taken by θ plays no role in the subsequent analysis.

To avoid technicalities, the diffusion tensor K is assumed to be piecewise constant on $\mathcal{T}_h$ and its restriction to an element $T \in \mathcal{T}_h$ is denoted by K_T. We will indicate by $\lambda_{m,T}$ and $\lambda_{M,T}$ respectively the minimum and the maximum eigenvalue of K on T. The degree of diffusion anisotropy on an element T is evaluated by the condition number of K_T, namely $\Delta_T = \frac{\lambda_{M,T}}{\lambda_{m,T}}$. Furthermore, the minimum value of $\tilde{\mu}$ on T is indicated by $\tilde{\mu}_{m,T}$. We assume that if $\tilde{\mu}_{m,T} = 0$, then $\|\mu\|_{L^\infty(T)} = \|\nabla\cdot\beta\|_{L^\infty(T)} = 0$.

3.3 A posteriori error analysis

3.3.1 Abstract setting

In this section we present the basic abstract framework for our a posteriori error estimates. The following result is directly inspired from the abstract framework introduced by Vohralík [102].

Lemma 3.1. *Let Z and Z_h be two vector spaces. Let A be a bilinear form defined on $Z^+ \times Z^+$ with $Z^+ := Z + Z_h$. Assume that A can be decomposed into the form $A = A_S + A_{SS}$ where A_S is symmetric and nonnegative on Z^+ and where A_{SS} is skew-symmetric on Z (but not necessarily on Z^+). Then, defining the semi-norm $|\cdot|_* := A_S(\cdot,\cdot)^{1/2}$, the following holds for all $u, s \in Z$ and $u_h \in Z_h$,*

$$|u - u_h|_* \le |s - u_h|_* + \sup_{\phi \in Z, |\phi|_* = 1} |A(u - u_h, \phi) + A_{SS}(u_h - s, \phi)|. \tag{3.14}$$

Proof. Let $u, s \in Z$ and $u_h \in Z_h$. Observe that if $u = s$, (3.14) obviously holds so that we may now suppose $u \neq s$. Suppose first that $|u - s|_* \le |u - u_h|_*$. Then,

$$\begin{aligned}
|u - u_h|_*^2 &= A(u - u_h, u - u_h) - A_{SS}(u - u_h, u - u_h) \\
&= A(u - u_h, u - s) + A(u - u_h, s - u_h) - A_{SS}(u - u_h, u - u_h) \\
&= A(u - u_h, u - s) + A_S(u - u_h, s - u_h) + A_{SS}(u - u_h, s - u_h) - A_{SS}(u - u_h, u - u_h) \\
&= A(u - u_h, u - s) + A_S(u - u_h, s - u_h) + A_{SS}(u - u_h, s - u) \\
&= A(u - u_h, u - s) + A_S(u - u_h, s - u_h) + A_{SS}(u_h - s, u - s),
\end{aligned}$$

where we have used $A_{SS}(u - s, u - s) = 0$ since $(u - s) \in Z$. Introducing $\phi_s = \frac{u-s}{|u-s|_*}$ and using the fact that for all $v, w \in Z^+$, $A_S(v, w) \le |v|_* |w|_*$ since A_S is symmetric and

nonnegative on Z^+ yields

$$|u-u_h|_*^2 \le |u-s|_* A(u-u_h,\phi_s) + |u-u_h|_*|s-u_h|_* + |u-s|_* A_{SS}(u_h-s,\phi_s). \quad (3.15)$$

Having hypothesized that $|u-s|_* \le |u-u_h|_*$, we infer

$$|u-u_h|_* \le |s-u_h|_* + |A(u-u_h,\phi_s) + A_{SS}(u_h-s,\phi_s)|, \quad (3.16)$$

whence (3.14) follows. Consider now the case $|u-u_h|_* \le |u-s|_*$. Since $A_{SS}(u-s,u-s)=0$,

$$\begin{aligned} |u-s|_*^2 &= A(u-s,u-s) = A(u-u_h,u-s) + A_S(u_h-s,u-s) + A_{SS}(u_h-s,u-s) \\ &\le |u-s|_* A(u-u_h,\phi_s) + |u_h-s|_*|u-s|_* + |u-s|_* A_{SS}(u_h-s,\phi_s). \end{aligned}$$

Thus,

$$|u-u_h|_* \le |u-s|_* \le A(u-u_h,\phi_s) + |s-u_h|_* + A_{SS}(u_h-s,\phi_s). \quad (3.17)$$

Combining the results we obtain (3.14). □

3.3.2 Pure diffusion

Let $\beta = 0$ and $\mu = 0$ in (3.1), i.e., we consider a diffusion problem with anisotropic and heterogeneous diffusivity:

$$\begin{cases} -\nabla\cdot(K\nabla u) = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \quad (3.18)$$

The bilinear form B defined by (3.7) becomes

$$B(v,w) = (K\nabla_h v, \nabla_h w)_{0,\Omega}, \quad (3.19)$$

while the definition of the (semi-)norm $\|\cdot\|_B$ involves only the diffusive contribution, i.e., $\|v\|_{B,T}^2 = (K\nabla_h v, \nabla_h v)_{0,T}$. The discrete problem is still (3.9) with the bilinear form B_h defined by

$$\begin{aligned} B_h(v,w) = (K\nabla_h v, \nabla_h w)_{0,\Omega} + \sum_{F\in\mathcal{F}_h} [&(\alpha h_F^{-1}\gamma_{K,F}[\![v]\!],[\![w]\!])_{0,F} \\ &- (n_F^t\{K\nabla_h v\}_\omega,[\![w]\!])_{0,F} - (n_F^t\{K\nabla_h w\}_\omega,[\![v]\!])_{0,F}]. \end{aligned} \quad (3.20)$$

Lemma 3.1 can be applied by letting $Z := V$, $Z_h := V_h$, $A = A_S := B$ and $A_{SS} := 0$. The semi-norm $|\cdot|_*$ coincides with $\|\cdot\|_B$. This yields

$$\|u - u_h\|_B \leq \inf_{s \in V} \|u_h - s\|_B + \sup_{\phi \in V, \|\phi\|_B = 1} |B(u - u_h, \phi)|. \tag{3.21}$$

We now proceed to estimate the second term in the right-hand side of (3.21). Let $\Pi_h : L^2(\Omega) \to V_h$ denote the L^2-orthogonal projection onto the vector space of piecewise constant functions on $\mathcal{T}_h$. It is well-known that for $v \in L^2(\Omega)$, $\Pi_h v$ coincides on each mesh element with the mean value of v on the corresponding element. The projector Π_h satisfies the following approximation properties: For all $T \in \mathcal{T}_h$ and for all $\phi \in H^1(T)$,

$$\|\phi - \Pi_h \phi\|_{0,T} \leq C_p^{\frac{1}{2}} h_T \|\nabla \phi\|_{0,T}, \leq C_p^{\frac{1}{2}} h_T \lambda_{m,T}^{-\frac{1}{2}} \|\phi\|_{B,T}, \tag{3.22}$$

$$\|\phi - \Pi_h \phi\|_{0,\partial T} \leq C_T^{\frac{1}{2}} h_T^{\frac{1}{2}} \|\nabla \phi\|_{0,T} \leq C_T^{\frac{1}{2}} h_T^{\frac{1}{2}} \lambda_{m,T}^{-\frac{1}{2}} \|\phi\|_{B,T}. \tag{3.23}$$

The constant C_p in the Poincaré-type inequality (3.22) can be bounded for each convex T by π^{-2}, see [19, 79], while it follows from [100] that the constant C_T in the trace inequality (3.23) is given by $C_T = 3d\rho_T$ with $\rho_T = h_T|\partial T|/|T|$ where $|\partial T|$ denotes the $(d-1)$-measure of ∂T and $|T|$ the d-measure of T; note that ρ_T is uniformly bounded owing to the shape-regularity of the mesh family. For all $T \in \mathcal{T}_h$, define on T the volumetric residual

$$R(u_h) = f + \nabla_h \cdot (K \nabla_h u_h), \tag{3.24}$$

and on ∂T the boundary residual such that for $F \subset \partial T$,

$$J_K(u_h)|_F = \bar{\omega}_{T,F} n_T^t [\![K \nabla_h u_h]\!] + \alpha h_F^{-1} \gamma_{K,F} [\![u_h]\!], \tag{3.25}$$

where

$$\bar{\omega}_{T,F} = 1 - \omega_{T,F}. \tag{3.26}$$

Note that $\bar{\omega}_{T,F} = 0$ on boundary faces.

Lemma 3.2. *The following holds:*

$$\sup_{\phi \in V, \|\phi\|_B = 1} |B(u - u_h, \phi)| \leq \left(\sum_{T \in \mathcal{T}_h} (\eta_T + \zeta_T)^2 \right)^{\frac{1}{2}}, \tag{3.27}$$

where the residual error indicator η_T is

$$\eta_T = C_p^{\frac{1}{2}} h_T \lambda_{m,T}^{-\frac{1}{2}} \|(I - \Pi_h) R(u_h)\|_{0,T}, \tag{3.28}$$

and the diffusive flux error indicator is

$$\zeta_T = C_T^{\frac{1}{2}} h_T^{\frac{1}{2}} \lambda_{m,T}^{-\frac{1}{2}} \|J_K(u_h)\|_{0,\partial T}. \tag{3.29}$$

Proof. Let $\phi \in V$ such that $\|\phi\|_B = 1$. Using $B(u,\phi) = (f,\phi)_{0,\Omega}$ and integrating by parts we obtain

$$B(u-u_h,\phi) = \sum_{T\in\mathcal{T}_h} (f + \nabla_h\cdot(K\nabla_h u_h),\phi)_{0,T} - \sum_{F\in\mathcal{F}_h^i} (n_F^t[\![K\nabla_h u_h]\!],\phi)_{0,F}$$

since $\phi \in V = H_0^1(\Omega)$. Testing the discrete equations with $\Pi_h\phi$ yields

$$\sum_{F\in\mathcal{F}_h} (\alpha h_F^{-1}\gamma_{K,F}[\![u_h]\!] - n_F^t\{K\nabla_h u_h\}_\omega, [\![\Pi_h\phi]\!])_{0,F} = (f,\Pi_h\phi)_{0,\Omega}.$$

On interior faces $F \in \mathcal{F}_h^i$, define the conjugate weighted average

$$\{v\}_{\bar\omega} = \omega_{T^+(F),F} v^- + \omega_{T^-(F),F} v^+,$$

so that $[\![vw]\!] = \{v\}_\omega [\![w]\!] + \{w\}_{\bar\omega}[\![v]\!]$ for any functions v and w which are (possibly) double-valued on F. Using this identity yields

$$\sum_{T\in\mathcal{T}_h} (\nabla_h\cdot(K\nabla_h u_h),\Pi_h\phi)_{0,T} = \sum_{F\in\mathcal{F}_h} (n_F^t\{K\nabla_h u_h\}_\omega, [\![\Pi_h\phi]\!])_{0,F} + \sum_{F\in\mathcal{F}_h^i} (n_F^t[\![K\nabla_h u_h]\!], \{\Pi_h\phi\}_{\bar\omega})_{0,F}.$$

Combining the above equations and using $[\![\phi]\!] = 0$ leads to

$$\begin{aligned} B(u-u_h,\phi) &= \sum_{T\in\mathcal{T}_h} (f + \nabla_h\cdot(K\nabla_h u_h),\phi-\Pi_h\phi)_{0,T} - \sum_{F\in\mathcal{F}_h} (\alpha h_F^{-1}\gamma_{K,F}[\![u_h]\!], [\![\phi-\Pi_h\phi]\!])_{0,F} \\ &\quad - \sum_{F\in\mathcal{F}_h^i} (n_F^t[\![K\nabla_h u_h]\!], \{\phi-\Pi_h\phi\}_{\bar\omega})_{0,F} \\ &= \sum_{T\in\mathcal{T}_h} (R(u_h),\phi-\Pi_h\phi)_{0,T} - \sum_{T\in\mathcal{T}_h}\sum_{F\subset\partial T} n_T\cdot n_F (J_K(u_h),\phi-\Pi_h\phi|_T)_{0,F}. \end{aligned}$$

The conclusion is straightforward using (3.22)–(3.23) and the fact that $\Pi_h(R(u_h))$ and $(\phi - \Pi_h\phi)$ are L^2-orthogonal on each $T \in \mathcal{T}_h$. □

Remark. Subtracting the mean value of $R(u_h)$ in the residual error estimator is possible because the discrete space contains piecewise constant functions. This is a feature of DG approximations, but not, for instance, of continuous finite element approximations.

Theorem 3.3. *Pick any $s_h \in V$ and define the non-conforming error indicator ι_T as*

$$\iota_T = \|u_h - s_h\|_{B,T}. \tag{3.30}$$

Then, the following holds

$$\|u - u_h\|_B \leq \left(\sum_{T \in \mathcal{T}_h} (\eta_T + \zeta_T)^2 \right)^{\frac{1}{2}} + \left(\sum_{T \in \mathcal{T}_h} \iota_T^2 \right)^{\frac{1}{2}}. \tag{3.31}$$

Proof. Direct consequence of Lemma 3.2 and of (3.21). □

We now investigate the local efficiency of the above error indicators η_T, ζ_T and ι_T. More precisely, we derive local upper bounds for these indicators in terms of the approximate error $u - u_h$, this time measured in the full energy norm, i.e. the energy semi-norm augmented by a term with jumps. Here, $x \lesssim y$ indicates the inequality $x \leq cy$ with positive c independent of the mesh and of the diffusion tensor. To simplify, the data f is assumed to be a polynomial; otherwise, the usual data oscillation term has to be added to the estimates. The following two propositions establish that the error indicators η_T and ζ_T are fully robust with respect to heterogeneities in the diffusion tensor, while the dependency on anisotropies remains local, i.e., only the square root of the diffusion condition numbers $\Delta_{\tilde{T}}$ on T and neighboring elements appears in the local lower bounds, but not the ratios of two diffusion tensor eigenvalues from different elements.

Proposition 3.4. *For all $T \in \mathcal{T}_h$,*

$$\eta_T \lesssim \Delta_T^{\frac{1}{2}} \|u - u_h\|_{B,T}. \tag{3.32}$$

Proof. Since $\|(I - \Pi_h)R(u_h)\|_{0,T} \leq \|R(u_h)\|_{0,T}$, we simply bound $\|R(u_h)\|_{0,T}$. To this purpose, we use the technique of element bubble functions introduced by Verfürth [96,97]; the arguments, which are fairly standard, are only briefly sketched. Let $T \in \mathcal{T}_h$, let b_T be a suitable local bubble function in T vanishing on ∂T and set $\nu_T = b_T R(u_h)$. Then,

$$\|R(u_h)\|_{0,T}^2 \lesssim (R(u_h), \nu_T)_{0,T} = (K\nabla_h(u - u_h), \nabla\nu_T)_{0,T} \lesssim \lambda_{M,T}^{\frac{1}{2}} h_T^{-1} \|u - u_h\|_{B,T} \|R(u_h)\|_{0,T}.$$

Hence,

$$\eta_T \lesssim h_T \lambda_{m,T}^{-\frac{1}{2}} \|R(u_h)\|_{0,T} \lesssim h_T \lambda_{m,T}^{-\frac{1}{2}} \lambda_{M,T}^{\frac{1}{2}} h_T^{-1} \|u - u_h\|_{B,T},$$

from which (3.32) follows. □

Proposition 3.5. *Let*

$$\|v\|_{B,*,\mathcal{F}}^2 = \sum_{F\in\mathcal{F}} \|\gamma_F^{\frac{1}{2}}[\![v]\!]\|_{0,F}^2 \qquad \forall v \in H^1(\mathcal{T}_h),$$

where we will either take $\mathcal{F} = \mathcal{F}_T$, $\mathcal{F} = \widetilde{\mathcal{F}}_T$ *or* $\mathcal{F} = \mathcal{F}_h$. *Then, for all* $T \in \mathcal{T}_h$,

$$\zeta_T \lesssim \Delta_T^{\frac{1}{2}} \left(\|u-u_h\|_{B,*,\mathcal{F}_T} + \sum_{\tilde{T}\in\mathcal{N}_T} \Delta_{\tilde{T}}^{\frac{1}{2}} \|u-u_h\|_{B,\tilde{T}} \right), \tag{3.33}$$

where $\mathcal{N}_T$ *is the set of elements sharing at least a face with the element* T.

Proof. Let $T \in \mathcal{T}_h$. Observe that

$$|\zeta_T| \lesssim \lambda_{m,T}^{-\frac{1}{2}} \sum_{F\in\mathcal{F}_T} \gamma_{K,F} h_F^{-\frac{1}{2}} \|[\![u_h]\!]\|_{0,F} + \lambda_{m,T}^{-\frac{1}{2}} h_T^{\frac{1}{2}} \sum_{F\in\mathcal{F}_T} \bar{\omega}_{T,F} \|n_F^t [\![K\nabla_h u_h]\!]\|_{0,F} \equiv X + Y,$$

and let us bound X and Y.
(i) Bound on X. There holds

$$X \lesssim \lambda_{m,T}^{-\frac{1}{2}} \sum_{F\in\mathcal{F}_T} \gamma_{K,F}^{\frac{1}{2}} \|\gamma_F^{\frac{1}{2}} [\![u_h]\!]\|_{0,F} \lesssim \Delta_T^{\frac{1}{2}} \|u-u_h\|_{B,*,\mathcal{F}_T},$$

since $\gamma_{K,F} \le n_F^t K_T n_F \le \lambda_{M,T}$.
(ii) Bound on Y. Let $F \in \mathcal{F}_T$. Using the technique of edge bubble functions introduced by Verfürth [96, 97], it is shown that

$$h_F^{\frac{1}{2}} \|n_F^t [\![K\nabla_h u_h]\!]\|_{0,F} \lesssim \sum_{T'\in\mathcal{T}(F)} \lambda_{M,T'}^{\frac{1}{2}} \|u-u_h\|_{B,T'}.$$

Hence,

$$\begin{aligned} Y &\lesssim \lambda_{m,T}^{-\frac{1}{2}} \sum_{F\in\mathcal{F}_T} \bar{\omega}_{T,F} \sum_{T'\in\mathcal{T}(F)} \lambda_{M,T'}^{\frac{1}{2}} \|u-u_h\|_{B,T'} \\ &\lesssim \Delta_T^{\frac{1}{2}} \sum_{F\in\mathcal{F}_T} \sum_{T'\in\mathcal{T}(F)} \lambda_{M,T}^{-\frac{1}{2}} \lambda_{m,T'}^{\frac{1}{2}} \bar{\omega}_{T,F} \Delta_{T'}^{\frac{1}{2}} \|u-u_h\|_{B,T'} \lesssim \Delta_T^{\frac{1}{2}} \sum_{F\in\mathcal{F}_T} \sum_{T'\in\mathcal{T}(F)} \Delta_{T'}^{\frac{1}{2}} \|u-u_h\|_{B,T'}, \end{aligned}$$

since

$$\lambda_{M,T}^{-\frac{1}{2}} \lambda_{m,T'}^{\frac{1}{2}} \bar{\omega}_{T,F} \le \frac{(n_F^t K_T n_F)^{\frac{1}{2}} (n_F^t K_{T'} n_F)^{\frac{1}{2}}}{(n_F^t K_T n_F) + (n_F^t K_{T'} n_F)} \le \frac{1}{2}.$$

The proof is complete. □

Remark. The local efficiency stated in Proposition 3.5 is given for the energy semi-norm augmented by the natural DG jump semi-norm $\|\cdot\|_{B,*,\mathcal{F}_h}$. Owing to the result of Ainsworth [5], global efficiency of ζ_T in the energy semi-norm $\|\cdot\|_B$ follows from (3.32)-(3.33) for sufficiently large stabilization parameters α in the case $d = 2$, $p = 1$, $K = Id$, and $\theta = 1$.

To analyze the local efficiency of the non-conforming error indicator ι_T, a particular choice must be made for $s_h \in V$. Presently, one of the state-of-the-art approaches consists in considering the so-called Oswald interpolate of the discrete solution u_h. For $v_h \in V_h$, its Oswald interpolate $\mathcal{I}_{Os}(v_h) \in V_h \cap V$ is defined by prescribing its values at the usual Lagrange interpolation nodes on each mesh element by taking the average of the values of v_h at the node,

$$\mathcal{I}_{Os}(v_h)(s) = \frac{1}{|\mathcal{T}_s|} \sum_{T \in \mathcal{T}_s} v_h|_T(s), \tag{3.34}$$

where $\mathcal{T}_s$ is the set of mesh elements that contain the node s and where $|\mathcal{T}_s|$ denotes the cardinal of that set. On boundary nodes, $\mathcal{I}_{Os}(v_h)(s)$ is set to zero. The Oswald interpolation operator $\mathcal{I}_{Os}$ yields the following local approximation properties [2, 62]: For all $v_h \in V_h$ and for all $T \in \mathcal{T}_h$,

$$\|v_h - \mathcal{I}_{Os}(v_h)\|_{0,T}^2 \leq C \sum_{F \in \widetilde{\mathcal{F}}_T} h_F \|[\![v_h]\!]\|_{0,F}^2, \tag{3.35}$$

$$\|\nabla_h(v_h - \mathcal{I}_{Os}(v_h))\|_{0,T}^2 \leq C \sum_{F \in \widetilde{\mathcal{F}}_T} h_F^{-1} \|[\![v_h]\!]\|_{0,F}^2, \tag{3.36}$$

where the constant C depends on the space dimension, the polynomial degree p used to construct the space V_h, and the shape-regularity parameter associated with the mesh $\mathcal{T}_h$; the dependency of the constant C on p has been recently explored in [24]. Setting $s_h := \mathcal{I}_{Os}(u_h)$ to evaluate ι_T, it is inferred using (3.36) that

$$\iota_T \lesssim \frac{\lambda_{M,T}^{\frac{1}{2}}}{\lambda_{m,\mathcal{R}_T}^{\frac{1}{2}}} \|u - u_h\|_{B,*,\widetilde{\mathcal{F}}_T} \tag{3.37}$$

where $\lambda_{m,\mathcal{R}_T} = \min_{T' \in \mathcal{R}_T} \lambda_{m,T'}$ and $\mathcal{R}_T = \{T' \in \mathcal{T}_h; T \cap T' \neq \emptyset\}$. Clearly, the above estimate is not robust with respect to heterogeneities and/or anisotropies in the diffusion tensor. In the isotropic case, the result can be improved by using weighted averages in (3.34) to define the nodal values of the Oswald interpolate. The weights depend on the diffusivity and a robust bound can be inferred on ι_T when evaluated with this modified

Oswald interpolate provided a monotonicity property of the diffusivity around vertices is assumed to hold; see [3, 22, 41].

To the authors' knowledge, no fully satisfactory result on a modified Oswald interpolation operator is yet available in the case of anisotropic diffusivity. We will not explore this issue further here. Finally, we point out that the error indicator ι_T can be readily sharpened by increasing the computational effort. Indeed, since any reconstructed function $s_h \in V$ can be chosen to evaluate it and since

$$\inf_{s\in V} \|u_h - s\|_{B,T} \le \|u_h - u\|_{B,T}, \tag{3.38}$$

the local efficiency of ι_T can be improved simply by solving more detailed local problems, and full robustness with respect to the diffusion tensor can eventually be achieved.

Remark. Using a triangle inequality, the flux error indicator ζ_T can be split into two contributions, one associated with the jump of the diffusive flux and the other associated with the jump of the discrete solution itself, and the latter can be regrouped with the nonconforming error indicator ι_T. By proceeding this way, the error upper bound is somewhat less sharp because a triangle inequality has been used, but the final form of the a posteriori error estimate takes a more familiar form.

3.3.3 Advection-diffusion-reaction

In this section we turn to the general case of an advection-diffusion-reaction problem. Our purpose is to extend the a posteriori error indicators derived in Lemma 3.2 and in Theorem 3.3 to this situation, with a particular emphasis on the robustness of the estimates in the high-Péclet regime in the sense of Verfürth [98]. The starting point is again the abstract estimate derived in Lemma 3.1 which is now applied with $Z := V$, $Z_h := V_h$,

$$A_S(v,w) = (K\nabla_h v, \nabla_h w)_{0,\Omega} + (\tilde{\mu} v, w)_{0,\Omega}, \tag{3.39}$$

$$A_{SS}(v,w) = (\beta\cdot\nabla_h v, w)_{0,\Omega} + \tfrac{1}{2}((\nabla\cdot\beta)v, w)_{0,\Omega}, \tag{3.40}$$

and $A = A_S + A_{SS} = B$ as defined by (3.7). Observe that A_S is symmetric and nonnegative on $Z + Z_h$, that $|\cdot|_*$ coincides with $\|\cdot\|_B$, and that A_{SS} is skew-symmetric on Z (but not on $Z + Z_h$). As a first step, we rewrite the quantity $B(u - u_h, \phi) + A_{SS}(u_h - s, \phi)$ in a more convenient form.

Lemma 3.6. *Let $s \in V$. For all $T \in \mathcal{T}_h$, define on T the volumetric residual*

$$R(u_h) = f + \nabla_h\cdot(K\nabla_h u_h) - \beta\cdot\nabla_h u_h - \mu u_h, \tag{3.41}$$

let $J_K(u_h)$ be defined on ∂T by (3.25), and let $J_\beta(u_h - s)$ be defined such that for $F \in \mathcal{F}_T$,

$$J_\beta(u_h - s)|_F = \langle \gamma_{\beta,F}[\![u_h - s]\!] + \beta{\cdot}n_F\{u_h - s\}\rangle_F, \tag{3.42}$$

where $\langle\cdot\rangle_F$ denotes the mean value over F. Then, for all $\phi \in V$,

$$B(u - u_h, \phi) + A_{SS}(u_h - s, \phi) = X_1 + X_2 + X_3, \tag{3.43}$$

with

$$X_1 = \sum_{T\in\mathcal{T}_h} ((I - \Pi_h)R(u_h), \phi - \Pi_h\phi)_{0,T}, \tag{3.44}$$

$$X_2 = -\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T} n_T{\cdot}n_F(J_K(u_h), \phi - \Pi_h\phi|_T)_{0,F}, \tag{3.45}$$

$$\begin{aligned} X_3 = &\sum_{T\in\mathcal{T}_h} [((I - \Pi_h)(\beta{\cdot}\nabla_h(u_h - s)), \phi - \Pi_h\phi)_{0,T} + \tfrac{1}{2}(\nabla{\cdot}\beta(u_h - s), \phi - 2\Pi_h\phi)_{0,T}] \\ &+ \sum_{F\in\mathcal{F}_h} (J_\beta(u_h - s), [\![\Pi_h\phi]\!])_{0,F}. \end{aligned} \tag{3.46}$$

Proof. Let $\phi \in V$. Using $B(u, \phi) = (f, \phi)_{0,\Omega}$ and integrating by parts, we infer

$$B(u - u_h, \phi) = \sum_{T\in\mathcal{T}_h} (R(u_h), \phi)_{0,T} - \sum_{F\in\mathcal{F}_h^i} (n_F^t[\![K\nabla_h u_h]\!], \phi)_{0,F}.$$

Testing the discrete equations with $\Pi_h\phi$ yields

$$\sum_{F\in\mathcal{F}_h} (\gamma_F[\![u_h]\!] - n_F^t\{K\nabla_h u_h\}_\omega + \beta{\cdot}n_F\{u_h\}, [\![\Pi_h\phi]\!])_{0,F} + ((\mu - \nabla{\cdot}\beta)u_h, \Pi_h\phi)_{0,\Omega} = (f, \Pi_h\phi)_{0,\Omega}.$$

Combining the two above equations and proceeding as in the proof of Lemma 3.2 for the diffusive term leads to

$$B(u - u_h, \phi) = X_1 + X_2 + \sum_{F\in\mathcal{F}_h} (\gamma_{\beta,F}[\![u_h]\!], [\![\Pi_h\phi]\!])_{0,F} - \sum_{F\in\mathcal{F}_h} (\beta{\cdot}n_F[\![u_h]\!], \{\Pi_h\phi\})_{0,F}.$$

Using the relation

$$\begin{aligned} &-\sum_{T\in\mathcal{T}_h} ((\nabla{\cdot}\beta)(u_h - s), \Pi_h\phi)_{0,T} - \sum_{T\in\mathcal{T}_h} (\beta{\cdot}\nabla_h(u_h - s), \Pi_h\phi)_{0,T} \\ &+ \sum_{F\in\mathcal{F}_h} (\beta{\cdot}n_F[\![u_h]\!], \{\Pi_h\phi\})_{0,F} + \sum_{F\in\mathcal{F}_h} (\beta{\cdot}n_F\{u_h - s\}, [\![\Pi_h\phi]\!])_{0,F} = 0, \end{aligned}$$

and adding $A_{SS}(u_h - s, \phi)$ as evaluated from (3.40), (3.43) is inferred. Note that the upwind related term $J_\beta(u_h - s)$ can be evaluated as a mean value over each face because it is tested against a piecewise constant function and that the mean value of $\beta{\cdot}\nabla_h(u_h - s)$ can be taken off on each element because it is tested against $\phi - \Pi_h\phi$. □

Remark. The idea of evaluating the upwind related term as a mean value over each face has been proposed by Vohralík [101]. Since for any function $\psi \in L^2(F)$, $\|\langle\psi\rangle_F\|_{0,F} \leq \|\psi\|_{0,F}$, this modification can only sharpen the a posteriori error estimate.

The next step is to control $\phi - \Pi_h\phi$ for $\phi \in V$ in terms of the energy norm $\|\phi\|_B$. To obtain bounds that behave satisfactorily when the Péclet number is large, a sharper version of inequalities (3.22)–(3.23) needs to be used. Observing that on all $T \in \mathcal{T}_h$, $\|\phi - \Pi_h\phi\|_{0,T} \leq \|\phi\|_{0,T}$ and letting

$$m_T = \min\left(C_p^{\frac{1}{2}} h_T \lambda_{m,T}^{-\frac{1}{2}}, \tilde{\mu}_{m,T}^{-\frac{1}{2}}\right), \tag{3.47}$$

the bound (3.22) can be sharpened as follows:

$$\|\phi - \Pi_h\phi\|_{0,T} \leq m_T \|\phi\|_{B,T}. \tag{3.48}$$

Furthermore, owing to the trace inequality

$$\forall v \in H^1(T), \quad \|v\|_{0,\partial T} \leq \rho_T^{\frac{1}{2}}[h_T^{-\frac{1}{2}}\|v\|_{0,T} + \|v\|_{0,T}^{\frac{1}{2}}\|\nabla v\|_{0,T}^{\frac{1}{2}}], \tag{3.49}$$

(see [28, 70] and section 3.6), (3.23) can be sharpened as follows:

$$\|\phi - \Pi_h\phi\|_{0,\partial T} \leq \rho_T^{\frac{1}{2}}[h_T^{-\frac{1}{2}} m_T + \lambda_{m,T}^{-\frac{1}{4}} m_T^{\frac{1}{2}}]\|\phi\|_{B,T} \leq \tilde{C}_{\mathcal{T}}^{\frac{1}{2}} \lambda_{m,T}^{-\frac{1}{4}} m_T^{\frac{1}{2}} \|\phi\|_{B,T}, \tag{3.50}$$

where we have set

$$\tilde{C}_{\mathcal{T}}^{\frac{1}{2}} = \rho_T^{\frac{1}{2}}(1 + C_p^{\frac{1}{4}}). \tag{3.51}$$

Estimate (3.50) will be used to bound the term X_2 introduced in Lemma 3.6. However, this estimate turns out not be sharp enough when bounding the last term in X_3. In this case, we will use the trace inequality

$$\forall \phi_h \in V_h, \quad \|\phi_h\|_{0,\partial T} \leq \rho_T^{\frac{1}{2}} h_T^{-\frac{1}{2}} \|\phi_h\|_{0,T}, \tag{3.52}$$

and we define for all $F \in \mathcal{F}_h$,

$$\tilde{m}_F^2 = \min\left(\max_{T' \in \mathcal{T}(F)} (C_{\mathcal{T}} h_{T'} \lambda_{m,T'}^{-1}), \max_{T' \in \mathcal{T}(F)} (\rho_{T'} h_{T'}^{-1} \tilde{\mu}_{m,T'}^{-1})\right), \tag{3.53}$$

recalling that $C_{T'} = 3d\rho_{T'}$. Finally, let $\kappa_{\mu,\beta,T} = \frac{1}{2}\|\nabla\cdot\beta\|_{L^\infty(T)} \tilde{\mu}_{m,T}^{-\frac{1}{2}}$. If $\tilde{\mu}_{m,T} = 0$, $\kappa_{\mu,\beta,T}$ should be evaluated as zero (recall that we have assumed $\|\nabla\cdot\beta\|_{L^\infty(T)} = 0$ in this case). To simplify the notation, we will use the convention $0/0 = 0$ in the sequel.

Lemma 3.7. *Let $s \in V$. The following holds*

$$\sup_{\phi\in V,\|\phi\|_B=1} |B(u-u_h,\phi)+A_{SS}(u_h-s,\phi)| \le \left(\sum_{T\in\mathcal{T}_h}(\eta_T+\zeta_T+\iota'_T)^2\right)^{\frac{1}{2}}, \tag{3.54}$$

where the residual error indicator η_T is

$$\eta_T = m_T\|(I-\Pi_h)R(u_h)\|_{0,T}, \tag{3.55}$$

the diffusive flux error indicator ζ_T is

$$\zeta_T = \tilde{C}_{\mathcal{T}}^{\frac{1}{2}}\lambda_{m,T}^{-\frac{1}{4}}m_T^{\frac{1}{2}}\|J_K(u_h)\|_{0,\partial T}, \tag{3.56}$$

and the non-conforming error indicator ι'_T is

$$\iota'_T = m_T\|(I-\Pi_h)(\beta\cdot\nabla_h(u_h-s))\|_{0,T}+\kappa_{\mu,\beta,T}\|u_h-s\|_{0,T}+\sum_{F\in\mathcal{F}_T}2\tilde{m}_F\|J_\beta(u_h-s)\|_{0,F}. \tag{3.57}$$

Proof. Let $\phi \in V$ such that $\|\phi\|_B = 1$. We bound the three terms X_1, X_2 and X_3 introduced in Lemma 3.6. Owing to (3.48) and (3.50), it is clear that

$$|X_1+X_2| \le \sum_{T\in\mathcal{T}_h}(\eta_T+\zeta_T)\|\phi\|_{B,T}.$$

Decompose X_3 into $X_3 = X_{3,1}+X_{3,2}$ where $X_{3,1}$ denotes the sum over elements and where $X_{3,2}$ denotes the sum over faces. Observing that $\|\phi-2\Pi_h\phi\|_{0,T} = \|\phi\|_{0,T}$ and using again (3.48), we obtain

$$|X_{3,1}| \le \sum_{T\in\mathcal{T}_h}(m_T\|(I-\Pi_h)(\beta\cdot\nabla_h(u_h-s))\|_{0,T}+\kappa_{\mu,\beta,T}\|u_h-s\|_{0,T})\|\phi\|_{B,T}.$$

To bound $X_{3,2}$, let $F \in \mathcal{F}_h$. On the one hand, owing to (3.23),

$$\begin{aligned}|(J_\beta(u_h-s),[\![\Pi_h\phi]\!])_{0,F}| &= |(J_\beta(u_h-s),[\![\Pi_h\phi-\phi]\!])_{0,F}|\\ &\le \sum_{T'\in\mathcal{T}(F)}|(J_\beta(u_h-s),\Pi_h\phi|_{T'}-\phi)_{0,F}|\\ &\le \|J_\beta(u_h-s)\|_{0,F}\max_{T'\in\mathcal{T}(F)}(C_{\mathcal{T}}^{\frac{1}{2}}h_{T'}^{\frac{1}{2}}\lambda_{m,T'}^{-\frac{1}{2}})\sum_{T'\in\mathcal{T}(F)}\|\phi\|_{B,T'}.\end{aligned}$$

On the other hand, owing to (3.52),

$$|(J_\beta(u_h - s), [\![\Pi_h \phi]\!])_{0,F}| \leq \sum_{T' \in \mathcal{T}(F)} |(J_\beta(u_h - s), \Pi_h \phi|_T)_{0,F}|$$
$$\leq \|J_\beta(u_h - s)\|_{0,F} \max_{T' \in \mathcal{T}(F)} (\rho_{T'}^{\frac{1}{2}} h_{T'}^{-\frac{1}{2}} \tilde{\mu}_{m,T'}^{-\frac{1}{2}}) \sum_{T' \in \mathcal{T}(F)} \|\phi\|_{B,T'}.$$

Hence,

$$|(J_\beta(u_h - s), [\![\Pi_h \phi]\!])_{0,F}| \leq \tilde{m}_F \|J_\beta(u_h - s)\|_{0,F} \sum_{T' \in \mathcal{T}(F)} \|\phi\|_{B,T'},$$

and therefore,

$$|X_{3,2}| \leq \sum_{T \in \mathcal{T}_h} \left(\sum_{F \in \mathcal{F}_T} 2\tilde{m}_F \|J_\beta(u_h - s)\|_{0,F} \right) \|\phi\|_{B,T}.$$

The conclusion is straightforward. □

Theorem 3.8. *Pick any $s_h \in V$ and define the non-conforming error indicator ι_T'' as*

$$\iota_T'' = \|u_h - s_h\|_{B,T}, \tag{3.58}$$

and let ι_T' be evaluated from (3.57) *using s_h. Then,*

$$\|u - u_h\|_B \leq \left(\sum_{T \in \mathcal{T}_h} (\eta_T + \zeta_T + \iota_T')^2 \right)^{\frac{1}{2}} + \left(\sum_{T \in \mathcal{T}_h} (\iota_T'')^2 \right)^{\frac{1}{2}}. \tag{3.59}$$

Proof. Apply Lemmata 3.1 and 3.7. □

Remark. The non-conforming error indicators ι_T' and ι_T'' can be regrouped into a single non-conforming error indicator ι_T by setting

$$\iota_T^2 = 4(\iota_T')^2 + 2(\iota_T'')^2. \tag{3.60}$$

Then, (3.59) becomes

$$\|u - u_h\|_B \leq \left(2 \sum_{T \in \mathcal{T}_h} (\eta_T + \zeta_T)^2 \right)^{\frac{1}{2}} + \left(\sum_{T \in \mathcal{T}_h} \iota_T^2 \right)^{\frac{1}{2}}, \tag{3.61}$$

which is less sharp but has a more familiar form.

We now investigate the local efficiency of the above error indicators η_T, ζ_T and ι_T. Here, $x \lesssim y$ indicates the inequality $x \leq cy$ with positive c independent of the mesh and of the parameters K, β, and μ. Again, the data f is assumed to be a polynomial; otherwise, the usual data oscillation term has to be added to the estimates. As in the pure diffusion case, we will not take advantage of the presence of the operator $(I - \Pi_h)$ in η_T and in the first term of ι'_T to derive the bounds below.

Proposition 3.9. *For all $T \in \mathcal{T}_h$,*

$$\eta_T \lesssim m_T[\lambda_{M,T}^{\frac{1}{2}} h_T^{-1} + \min(\alpha_{1,T}, \alpha_{2,T})]\|u - u_h\|_{B,T}, \tag{3.62}$$

where

$$\alpha_{1,T} = \frac{\|\mu\|_{L^\infty(T)}}{\tilde{\mu}_{m,T}^{\frac{1}{2}}} + \frac{\|\beta\|_{L^\infty(T)}}{\lambda_{m,T}^{\frac{1}{2}}}, \qquad \alpha_{2,T} = \frac{\|\mu - \nabla\cdot\beta\|_{L^\infty(T)} + \|\beta\|_{L^\infty(T)} h_T^{-1}}{\tilde{\mu}_{m,T}^{\frac{1}{2}}}.$$

Proof. Let $T \in \mathcal{T}_h$, let b_T be a suitable local bubble function in T vanishing on ∂T and set $\nu_T = b_T R(u_h)$. Then,

$$\begin{aligned}
\|R(u_h)\|_{0,T}^2 &\lesssim (R(u_h), \nu_T)_{0,T} = (K\nabla_h(u - u_h), \nabla_h \nu_T)_{0,T} + (\mu(u - u_h), \nu_T)_{0,T} \\
&\quad + (\beta\cdot\nabla_h(u - u_h), \nu_T)_{0,T} \\
&\lesssim \lambda_{M,T}^{\frac{1}{2}} h_T^{-1} \|u - u_h\|_{B,T} \|R(u_h)\|_{0,T} + \min\left(\alpha_{1,T}, \alpha_{2,T}\right) \|u - u_h\|_{B,T} \|R(u_h)\|_{0,T},
\end{aligned}$$

where the min is obtained by integrating by parts or not the advective derivative. The conclusion is straightforward. □

Proposition 3.10. *For all $T \in \mathcal{T}_h$,*

$$\zeta_T \lesssim \Delta_T^{\frac{1}{2}} \|u - u_h\|_{B,*,\mathcal{F}_T} + \Delta_T^{\frac{1}{2}} \lambda_{m,T}^{\frac{1}{4}} m_T^{\frac{1}{2}} \sum_{\tilde{T} \in \mathcal{N}_T} m_{\tilde{T}}^{-\frac{1}{2}} \lambda_{m,\tilde{T}}^{-\frac{1}{4}} \left(m_{\tilde{T}} \alpha_{1,\tilde{T}} + \Delta_{\tilde{T}}^{\frac{1}{2}} \right) \|u - u_h\|_{B,\tilde{T}}. \tag{3.63}$$

Proof. Let $T \in \mathcal{T}_h$. Observe that

$$|\zeta_T| \lesssim \lambda_{m,T}^{-\frac{1}{4}} m_T^{\frac{1}{2}} \sum_{F \in \mathcal{F}_T} \gamma_{K,F} h_F^{-1} \|[\![u_h]\!]\|_{0,F} + \lambda_{m,T}^{-\frac{1}{4}} m_T^{\frac{1}{2}} \sum_{F \in \mathcal{F}_T} \bar{\omega}_{T,F} \|n_F^t [\![K\nabla_h u_h]\!]\|_{0,F} \equiv X + Y,$$

and let us bound X and Y by the right-hand side of (3.63).
(i) Bound on X. Owing to the definition of $\gamma_{K,F}$,

$$\begin{aligned} X &\lesssim \lambda_{m,T}^{-\frac{1}{4}}\lambda_{M,T}^{\frac{1}{2}}h_T^{-\frac{1}{2}}m_T^{\frac{1}{2}}\sum_{F\in\mathcal{F}_T}\gamma_{K,F}{}^{\frac{1}{2}}h_F^{-\frac{1}{2}}\|[\![u_h]\!]\|_{0,F} \\ &\lesssim \Delta_T^{\frac{1}{2}}\lambda_{m,T}^{\frac{1}{4}}h_T^{-\frac{1}{2}}m_T^{\frac{1}{2}}\|u-u_h\|_{B,*,\mathcal{F}_T}. \end{aligned}$$

Owing to the obvious bound $h_T^{-\frac{1}{2}} \lesssim m_T^{-\frac{1}{2}}\lambda_{m,T}^{-\frac{1}{4}}$, it is inferred that X is bounded by the first term on the right-hand side of (3.63).
(ii) Bound on Y. Let $F \in \mathcal{F}_T$. Following the ideas of Verfürth [98], let b_F be a suitable bubble function with support in F and let ℓ_F be the lifting of $(n_F^t[\![K\nabla_h u_h]\!])b_F$ in $\mathcal{T}(F)$ with cut-off parameter

$$\theta_{T'} = m_{T'}C_p^{-\frac{1}{2}}h_{T'}^{-1}\lambda_{m,T'}^{\frac{1}{2}} \le 1,$$

on each $T' \in \mathcal{T}(F)$. Then,

$$\begin{aligned} \|n_F^t[\![K\nabla_h u_h]\!]\|_{0,F}^2 &\lesssim (n_F^t[\![K\nabla_h u_h]\!],\ell_F)_{0,F}, \\ \|\ell_F\|_{0,T'} &\lesssim h_{T'}^{\frac{1}{2}}\theta_{T'}^{\frac{1}{2}}\|n_F^t[\![K\nabla_h u_h]\!]\|_{0,F} \lesssim m_{T'}^{\frac{1}{2}}\lambda_{m,T'}^{\frac{1}{4}}\|n_F^t[\![K\nabla_h u_h]\!]\|_{0,F}, \\ \|\nabla\ell_F\|_{0,T'} &\lesssim h_{T'}^{-\frac{1}{2}}\theta_{T'}^{-\frac{1}{2}}\|n_F^t[\![K\nabla_h u_h]\!]\|_{0,F} \lesssim m_{T'}^{-\frac{1}{2}}\lambda_{m,T'}^{-\frac{1}{4}}\|n_F^t[\![K\nabla_h u_h]\!]\|_{0,F}. \end{aligned}$$

Observe that

$$B(u-u_h,\ell_F) = (R(u_h),\ell_F)_{0,\mathcal{T}(\mathcal{F})} + (n_F^t[\![K\nabla_h u_h]\!],\ell_F)_{0,F},$$

and that

$$|B(u-u_h,\ell_F)| \lesssim \sum_{T'\in\mathcal{T}(F)}(\lambda_{M,T'}^{\frac{1}{2}}m_{T'}^{-\frac{1}{2}}\lambda_{m,T'}^{-\frac{1}{4}} + m_{T'}^{\frac{1}{2}}\lambda_{m,T'}^{\frac{1}{4}}\alpha_{1,T'})\|u-u_h\|_{B,T'}\|n_F^t[\![K\nabla_h u_h]\!]\|_{0,F}.$$

Furthermore, since

$$\begin{aligned} |(R(u_h),\ell_F)_{0,\mathcal{T}(\mathcal{F})}| &\le \sum_{T'\in\mathcal{T}(F)}\|R(u_h)\|_{0,T'}\|\ell_F\|_{0,T'} \\ &\lesssim \sum_{T'\in\mathcal{T}(F)}[\lambda_{M,T'}^{\frac{1}{2}}h_{T'}^{-1} + \min(\alpha_{1,T'},\alpha_{2,T'})]\|u-u_h\|_{B,T'}\|\ell_F\|_{0,T'} \\ &\lesssim \sum_{T'\in\mathcal{T}(F)}[\lambda_{M,T'}^{\frac{1}{2}}h_{T'}^{-1} + \alpha_{1,T'}]m_{T'}^{\frac{1}{2}}\lambda_{m,T'}^{\frac{1}{4}}\|u-u_h\|_{B,T'}\|n_F^t[\![K\nabla_h u_h]\!]\|_{0,F}, \end{aligned}$$

and since $h_{T'}^{-1}m_{T'}^{\frac12}\lambda_{m,T'}^{\frac14} \le m_{T'}^{-\frac12}\lambda_{m,T'}^{-\frac14}$, it is inferred that $|(R(u_h),\ell_F)_{0,\mathcal{T}(\mathcal{F})}|$ can be bounded as $|B(u-u_h,\ell_F)|$, whence

$$\|n_F^t[\![K\nabla_h u_h]\!]\|_{0,F} \lesssim \sum_{T'\in\mathcal{T}(F)} (\lambda_{M,T'}^{\frac12}m_{T'}^{-\frac12}\lambda_{m,T'}^{-\frac14} + m_{T'}^{\frac12}\lambda_{m,T'}^{\frac14}\alpha_{1,T'})\|u-u_h\|_{B,T'}.$$

As a result,

$$\begin{aligned}
Y &\lesssim \lambda_{m,T}^{-\frac14}m_T^{\frac12}\sum_{F\in\mathcal{F}_T}\sum_{T'\in\mathcal{T}(F)} \bar\omega_{T,F}(\lambda_{M,T'}^{\frac12}m_{T'}^{-\frac12}\lambda_{m,T'}^{-\frac14} + m_{T'}^{\frac12}\lambda_{m,T'}^{\frac14}\alpha_{1,T'})\|u-u_h\|_{B,T'}\\
&\lesssim \Delta_T^{\frac12}\lambda_{m,T}^{\frac14}m_T^{\frac12}\sum_{F\in\mathcal{F}_T}\sum_{T'\in\mathcal{T}(F)} (\lambda_{M,T}^{-\frac12}\lambda_{m,T'}^{\frac12}\bar\omega_{T,F}(\Delta_{T'}^{\frac12} + m_{T'}\alpha_{1,T'})m_{T'}^{-\frac12}\lambda_{m,T'}^{-\frac14}\|u-u_h\|_{B,T'}\\
&\lesssim \Delta_T^{\frac12}\lambda_{m,T}^{\frac14}m_T^{\frac12}\sum_{\tilde T\in\mathcal{N}_T} (\Delta_{\tilde T}^{\frac12} + m_{\tilde T}\alpha_{1,\tilde T})m_{\tilde T}^{-\frac12}\lambda_{m,\tilde T}^{-\frac14}\|u-u_h\|_{B,\tilde T}.
\end{aligned}$$

The conclusion is straightforward. □

Finally, we investigate the local efficiency of the non-conforming error estimator ι_T. To this purpose, we pick $s_h = \mathcal{I}_{Os}(u_h)$. As discussed at the end of §3.3.2, a modified Oswald interpolation operator can be considered in the case of isotropic and heterogeneous diffusivity with a monotonicity property around vertices to sharpen the result.

Proposition 3.11. *Set $s_h = \mathcal{I}_{Os}(u_h)$. Let $T\in\mathcal{T}_h$. Define $c_{\beta,\widetilde{\mathcal{F}}_T} = \min_{F\in\widetilde{\mathcal{F}}_T}\gamma_{\beta,F}$. Then,*

$$\begin{aligned}
\iota_T \lesssim{}& \left(\lambda_{M,T}^{\frac12}h_T^{-1} + \|\tilde\mu\|_{L^\infty(T)}^{\frac12} + m_T\|\beta\|_{L^\infty(T)}h_T^{-1} + \kappa_{\mu,\beta,T} + \sum_{F\in\mathcal{F}_T}\tilde m_F\|\beta\cdot n_F\|_{L^\infty(F)}h_T^{-\frac12}\right)\\
&\times\min\left(\frac{h_T}{\lambda_{m,\mathcal{R}_T}^{\frac12}}, \frac{h_T^{\frac12}}{c_{\beta,\widetilde{\mathcal{F}}_T}^{\frac12}}\right)\|u-u_h\|_{B,*,\widetilde{\mathcal{F}}_T}.
\end{aligned}\tag{3.64}$$

Proof. Let $T\in\mathcal{T}_h$. Observe first that

$$\sum_{F\in\widetilde{\mathcal{F}}_T}\|[\![u_h]\!]\|_{0,F} \le \min\left(\frac{h_T}{\lambda_{m,\mathcal{R}_T}^{\frac12}}, \frac{h_T^{\frac12}}{c_{\beta,\widetilde{\mathcal{F}}_T}^{\frac12}}\right)h_T^{-\frac12}\|u-u_h\|_{B,*,\widetilde{\mathcal{F}}_T}.$$

Hence, using (3.35)–(3.36),

$$\|u_h - s_h\|_{B,T} \lesssim \left(\lambda_{M,T}^{\frac{1}{2}} h_T^{-1} + \|\tilde{\mu}\|_{L^\infty(T)}^{\frac{1}{2}}\right) \min\left(\frac{h_T}{\lambda_{m,\mathcal{R}_T}^{\frac{1}{2}}}, \frac{h_T^{\frac{1}{2}}}{c_{\beta,\widetilde{\mathcal{F}}_T}^{\frac{1}{2}}}\right) \|u - u_h\|_{B,*,\widetilde{\mathcal{F}}_T},$$

where $\lambda_{m,\mathcal{R}_T} = \min_{T' \in \mathcal{R}_T} \lambda_{m,T'}$ and $\mathcal{R}_T = \{T' \in \mathcal{T}_h; T \cap T' \neq \emptyset\}$. Furthermore, still using (3.35)–(3.36), the first two terms in ι'_T (see (3.57)) are bounded by

$$\left(m_T \|\beta\|_{L^\infty(T)} h_T^{-1} + \kappa_{\mu,\beta,T}\right) \min\left(\frac{h_T}{\lambda_{m,\mathcal{R}_T}^{\frac{1}{2}}}, \frac{h_T^{\frac{1}{2}}}{c_{\beta,\widetilde{\mathcal{F}}_T}^{\frac{1}{2}}}\right) \|u - u_h\|_{B,*,\widetilde{\mathcal{F}}_T},$$

and it remains to bound the last term, namely $\sum_{F \in \mathcal{F}_T} 2\tilde{m}_F \|J_\beta(u_h - s_h)\|_{0,F}$. For all $T \in \mathcal{T}_h$, it can be shown that $\forall v_h \in V_h$,

$$\|\{u_h - \mathcal{I}_{Os}(u_h)\}\|_{0,\partial T} \lesssim \sum_{F \in \widetilde{\mathcal{F}}_T} \|[\![u_h]\!]\|_{0,F},$$

whence the conclusion is straightforward. □

To illustrate by a simple example, assume that β and μ are of order unity, that β is solenoidal (or that its divergence is uniformly bounded by $\tilde{\mu}$ locally), and that the diffusion is homogeneous and isotropic, i.e., $K = \epsilon I_d$ with real parameter $0 < \epsilon \leq 1$ and where I_d denotes the identity matrix in $\mathbb{R}^d$. Then, $m_T = \min(h_T \epsilon^{-\frac{1}{2}}, 1)$, $\alpha_{1,T} = 1 + \epsilon^{-\frac{1}{2}}$, $\alpha_{2,T} = 1 + h_T^{-1}$, and it is readily verified that all the constants appearing in the local estimates for η_T, ζ_T, and ι_T are bounded by $(1 + \epsilon^{-\frac{1}{2}} \min(h_T \epsilon^{-\frac{1}{2}}, 1))$, which corresponds to the result derived in [98] for continuous finite elements with vanishing, isotropic, and homogeneous diffusion.

3.4 Numerical results

In this section, the present a posteriori error estimators are assessed on two test cases. The first one is a pure diffusion problem with heterogeneous isotropic diffusion; its aim is to verify numerically the sharpness of the diffusion flux error indicator ζ_T when evaluated with the proper weights. The second test case is an advection–diffusion-reaction problem with homogeneous diffusion; its aim is to verify the behavior of the a posteriori error estimates in the low- and high-Péclet number regimes. We have always taken $\theta = 1$ in (3.10),

while α in the definition of γ_F has been taken equal to 4. The corresponding DG method is the so-called Symmetric Weighted Interior Penalty method analyzed recently in [51]. Moreover, we have set $p = 1$, i.e., used piecewise linears. In all cases, the non-conforming error indicators have been evaluated using the standard Oswald interpolate of the discrete solution; see (3.34).

3.4.1 Heterogeneous diffusion

We consider the following test problem proposed in [88]. The domain $\Omega = (-1,1)\times(-1,1)$ is split into four subregions: $\Omega_1 = (0,1)\times(0,1)$, $\Omega_2 = (-1,0)\times(0,1)$, $\Omega_3 = (-1,0)\times(-1,0)$, and $\Omega_4 = (0,1)\times(-1,0)$. The source term f is zero. The diffusion tensor is isotropic, i.e., of the form $K|_{\Omega_i} = \epsilon_i I$ with constant value within each subregion for $i \in \{1,2,3,4\}$. Letting $\epsilon_1 = \epsilon_3 = 100$ and $\epsilon_2 = \epsilon_4 = 1$, the exact solution written in polar coordinates is

$$u|_{\Omega_i} = r^\alpha \left(a_i \sin(\alpha\theta) + b_i \cos(\alpha\theta)\right), \tag{3.65}$$

with $\alpha = 0.12690207$ and

$$\begin{aligned} a_1 &= 0.100000000 & b_1 &= 1.000000000,\\ a_2 &= -9.603960396 & b_2 &= 2.960396040,\\ a_3 &= -0.480354867 & b_3 &= -0.882756593,\\ a_4 &= 7.701564882 & b_4 &= -6.456461752. \end{aligned}$$

Non-homogeneous Dirichlet boundary conditions as given by (3.65) are enforced on $\partial\Omega$. The exact solution possesses a singularity at the origin, and its regularity depends on the constant α, namely $u \in H^\alpha(\Omega)$; see [80] for further regularity results for this type of problems. The expected convergence order of the error in the L^2-norm is 2α, while the expected convergence order in the energy semi-norm is α. Table 3.1 presents the results on a series of quasi-uniform unstructured triangulations with N mesh elements. All the meshes are compatible with the above partition of the domain Ω. The last line of the table displays the convergence orders evaluated on the last two meshes. The convergence orders for the error both in the L^2-norm and in the energy semi-norm are in good agreement with the theoretical predictions. The same conclusion is reached for the a posteriori error estimators based on ζ_T and ι_T (observe that in the present case, $\eta_T = 0$ because $f = 0$ and $p = 1$). The column labelled "est" reports the total a posteriori error estimator derived in Theorem 3.3, and the column labelled "eff" reports the effectivity index of the estimator, namely the ratio of the a posteriori error estimator to the actual approximation error. The

effectivity index is about 7 on all meshes. Notice that all the constants in the estimators are explicitly evaluated. To compare, using the more conventional DG method based on arithmetic averages (i.e., weights equal to $\frac{1}{2}$ on all faces) and a penalty term $\gamma_{K,F}$ equal to the arithmetic mean of the normal diffusivities on each face, the effectivity indices are about 5 times larger.

N	$\|u-u_h\|_{0,\Omega}$	$\|u-u_h\|_B$	$(\sum_{T\in\mathcal{T}_h}\zeta_T^2)^{\frac{1}{2}}$	$(\sum_{T\in\mathcal{T}_h}\iota_T^2)^{\frac{1}{2}}$	est.	eff.
112	8.48e-2	3.27	15.62	11.84	27.45	8.4
448	7.16e-2	3.11	10.92	11.35	22.26	7.2
1792	6.19e-2	2.93	9.89	10.82	20.72	7.1
7168	5.37e-2	2.75	9.16	10.27	19.43	7.1
order	0.21	0.09	0.11	0.08	0.09	–

Table 3.1: Heterogeneous diffusion with parameter $\alpha = 0.13$

We have also examined a similar test case with a less singular solution corresponding to milder contrasts in the diffusion, namely $\epsilon_1 = \epsilon_3 = 5$ and $\epsilon_2 = \epsilon_4 = 1$. In this case, the exact solution is still given by (3.65) with $\alpha = 0.53544095$ and

$$\begin{aligned} a_1 &= 0.44721360 & b_1 &= 1.00000000, \\ a_2 &= -0.74535599 & b_2 &= 2.33333333, \\ a_3 &= -0.94411759 & b_3 &= 0.55555556, \\ a_4 &= -2.40170264 & b_4 &= -0.48148148. \end{aligned}$$

Table 3.2 presents the results. The conclusions are similar to those reached with the previous test case. The effectivity index is approximately equal to 7 (except on the coarsest mesh), and thus takes comparable values to those in the previous test case, confirming the robustness of the estimates with respect to diffusion heterogeneities.

N	$\|u-u_h\|_{0,\Omega}$	$\|u-u_h\|_B$	$(\sum_{T\in\mathcal{T}_h}\zeta_T^2)^{\frac{1}{2}}$	$(\sum_{T\in\mathcal{T}_h}\iota_T^2)^{\frac{1}{2}}$	est.	eff.
112	2.66e-2	6.11e-1	4.82	8.70e-1	5.69	9.32
448	1.13e-2	4.28e-1	2.49	6.09e-1	3.10	7.23
1792	4.98e-3	2.97e-1	1.66	4.23e-1	2.08	7.00
7168	2.26e-3	2.06e-1	1.13	2.92e-1	1.42	6.90
order	1.14	0.53	0.56	0.53	0.55	–

Table 3.2: Heterogeneous diffusion with parameter $\alpha = 0.54$

We conclude this section by an example on how the error estimator can be used to adapt the mesh. We consider the test case with parameter $\alpha = 0.54$. Starting from the quasi-uniform mesh with $N = 112$ considered previously, the adaptive mesh refinement procedure flags 5% of the mesh elements yielding the largest error indicators. Results are reported in Table 3.3. The efficiency of the procedure can be seen for instance by observing that the error in the energy semi-norm is approximately 0.29 on an adaptive mesh with $N = 288$ elements, while $N = 1792$ elements are needed in a quasi-uniform mesh to achieve the same target. Figure 3.1 presents two meshes obtained with the adaptive refinement procedure, one with 148 elements and one with 200 elements. We see that the adaptive refinement correctly aims at capturing the singularity at the origin.

N	$\|u-u_h\|_B$	$(\sum_{T\in\mathcal{T}_h}\zeta_T^2)^{\frac{1}{2}}$	$(\sum_{T\in\mathcal{T}_h}\iota_T^2)^{\frac{1}{2}}$	eff.
112	6.11e-1	5.69	8.70e-1	9.3
148	4.58e-1	2.53	6.17e-1	5.5
200	3.51e-1	2.20	4.40e-1	6.3
288	2.86e-1	2.00	3.17e-1	7.0
394	2.69e-1	1.72	3.13e-1	6.4

Table 3.3: Error as a function of mesh elements on adaptive meshes

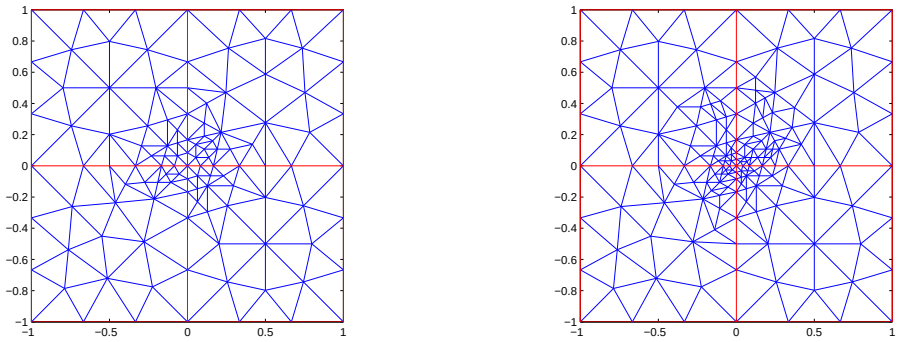


Figure 3.1: Two adaptive meshes derived from the error estimator: 148 elements (left) and 200 elements (right)

3.4.2 Advection-diffusion-reaction

Consider the domain $\Omega = (0,1) \times (0,1)$, the advection field $\beta = (1,0)^t$, the reaction coefficient $\mu = 1$, and an isotropic homogeneous diffusion tensor $K = \epsilon I$. We run tests with $\epsilon = 1$ and $\epsilon = 10^{-4}$ to examine the difference between dominant diffusion and dominant advection regimes. Since the diffusion is homogeneous and isotropic, the SWIP method coincides with the more conventional Interior Penalty DG method. The source term f is designed so that the exact solution is

$$u(x,y) = 0.5x(1-x)y(1-y)\left(1 - \tanh\left(\frac{0.5-x}{\gamma}\right)\right). \tag{3.66}$$

Here, the parameter $\gamma = 0.05$ controls the thickness of the internal layer at $x = 0.5$. Homogeneous Dirichlet boundary conditions are enforced.

In Table 3.4 we present the results for the dominant diffusion regime. The estimator and the error converge at the same order, and the effectivity index is comparable with that obtained for a pure diffusion problem. The dominant contribution to the total a posteriori error estimate is the diffusive flux error indicator. When the advection becomes dominant (Table 3.5), the main contribution is the non-conforming error indicator ι'_T and, marginally, the residual error indicator (which converges to second-order owing to the subtraction of the elementwise mean-value of the residue). Owing to the appropriate use of cut-off functions, the effectivity index is only twenty times larger than in the dominant diffusion regime. The last line of Table 3.5 reports the convergence orders evaluated using the last two meshes. It can be observed that the last mesh is sufficiently fine, leading to first-order convergence

of $\|u-u_h\|_B$, whereas this quantity converges to order 1.5 on coarser meshes where the L^2-contribution dominates. Finally, adaptive meshes can be generated using the above error indicators (not shown). As expected, the adaptive refinement occurs in the vicinity of the internal layer.

N	$\|u-u_h\|_B$	$(\sum_{T\in\mathcal{T}_h}\eta_T^2)^{\frac{1}{2}}$	$(\sum_{T\in\mathcal{T}_h}\zeta_T^2)^{\frac{1}{2}}$	$(\sum_{T\in\mathcal{T}_h}\iota_T'^2)^{\frac{1}{2}}$	$(\sum_{T\in\mathcal{T}_h}\iota_T''^2)^{\frac{1}{2}}$	est.	eff.
256	4.39e-2	3.48e-2	3.13e-1	1.13e-2	1.90e-2	3.74e-1	8.5
1024	2.28e-2	1.03e-2	1.74e-1	2.61e-3	1.11e-2	1.96e-1	8.6
4096	1.15e-2	2.68e-3	9.16e-2	5.51e-4	5.45e-3	9.71e-2	8.4
16384	5.71e-3	6.76e-4	4.65e-2	1.28e-4	2.61e-3	4.97e-2	8.7
order	1.01	1.98	0.99	2.10	1.07	0.97	–

Table 3.4: Advection-diffusion with $\epsilon = 1$

N	$\|u-u_h\|_B$	$(\sum_{T\in\mathcal{T}_h}\eta_T^2)^{\frac{1}{2}}$	$(\sum_{T\in\mathcal{T}_h}\zeta_T^2)^{\frac{1}{2}}$	$(\sum_{T\in\mathcal{T}_h}\iota_T'^2)^{\frac{1}{2}}$	$(\sum_{T\in\mathcal{T}_h}\iota_T''^2)^{\frac{1}{2}}$	est.	eff.
256	7.85e-4	3.85e-2	2.43e-3	6.70e-2	9.33e-4	1.05e-1	134
1024	2.98e-4	2.00e-2	1.88e-3	3.48e-2	2.47e-4	5.56e-2	186
4096	1.32e-4	8.40e-3	1.24e-3	1.74e-2	8.80e-5	2.67e-2	201
16384	6.40e-5	2.18e-3	6.23e-4	8.53e-3	3.90e-5	1.13e-2	177
order	1.06	1.95	0.99	1.03	1.17	1.25	–

Table 3.5: Advection-diffusion with $\epsilon = 10^{-4}$

3.5 Conclusions

In this work, we have proposed and analyzed a posteriori energy-norm error estimates for weighted interior penalty DG approximations to advection-diffusion-reaction equations with heterogeneous and anisotropic diffusion. All the constants in the error upper bounds have been specified, so that the present estimates can be used for actual control over the error in practical simulations. Local lower error bounds in which all the dependencies on model parameters are explicitly stated, have been derived as well. In the case of pure diffusion, full robustness is achieved with respect to diffusion heterogeneities owing to the use of suitable diffusion-dependent weights to formulate the consistency terms in the DG method. This feature has been verified numerically and stands in contrast to the results

obtained with more conventional interior penalty DG approximations. Furthermore, diffusion anisotropies enter the lower error bounds only through the square root of the condition number of the diffusion tensor on a given mesh cell and its neighbors. Current state-of-the-art results have been used to evaluate the non-conforming error estimators through the so-called Oswald interpolate; further work in this direction is needed to investigate the robustness with respect to diffusion heterogeneities and anisotropies. In the presence of advection, we have shown, in the spirit of the work of Verfürth for continuous finite element approximations with streamline diffusion stabilization, that the lower error bounds can be written with constants involving a cut-off for the ratio of local mesh size to the reciprocal of the square root of the lowest local eignevalue of the diffusion tensor.

3.6 Appendix: Trace inequality

The following Lemma is a slight variation of that found in the article by Monk and Süli [70] and that of Carstensen and Funken [28].

Lemma 3.12. *Set $d \geq 2$. For all $T \in \mathcal{T}_h$ and $v \in H^1(T)$,*

$$\|v\|_{0,\partial T} \leq \rho_T^{\frac{1}{2}} \|v\|_{0,T}^{\frac{1}{2}} \left(h_T^{-\frac{1}{2}} \|v\|_{0,T}^{\frac{1}{2}} + \|\nabla v\|_{0,T}^{\frac{1}{2}} \right) \tag{3.67}$$

with

$$\rho_T = \frac{|\partial T| h_T}{|T|} \tag{3.68}$$

where $|T|$ denotes the d-measure of T and $|\partial T|$ the $(d-1)$-measure of ∂T.

Proof. For $0 \leq i \leq d$, let p_i be a function defined on $T \in \mathcal{T}_h$ by

$$p_i = \frac{|F_i|}{d|T|}(x - a_i)$$

where F_i is the face opposite to node a_i. Note that the normal component is equal to 1 on F_i and vanishes on F_j, $j \neq i$. Furthermore,

$$\|p_i\|_{L^\infty(T)} \leq \frac{|F_i| h_T}{d|T|} \qquad \text{and} \qquad \nabla \cdot p_i = \frac{|F_i|}{|T|}.$$

Using the divergence theorem and the Cauchy-Schwarz inequality we have

$$\begin{aligned}\|u\|_{0,\partial T}^2 &\leq \sum_{i=0}^{d}(u^2, p_i \cdot n_T)_{0,\partial T} \leq \sum_{i=0}^{d}(\nabla\cdot(u^2 p_i), 1)_{0,T} \\ &\leq \sum_{i=0}^{d}(2up_i\nabla u + u^2\nabla\cdot p_i, 1)_{0,T} \leq \sum_{i=0}^{d}(2\|u\|_{0,T}\|\nabla u\|_{0,T}\|p_i\|_{L^\infty(T)} + \|u\|_{0,T}^2\frac{|F_i|}{|T|}) \\ &\leq \sum_{i=0}^{d}\frac{|F_i|h_T}{|T|}\|u\|_{0,T}\left(\frac{2}{d}h_T^{-1}\|u\|_{0,T} + \|\nabla u\|_{0,T}\right).\end{aligned}$$

Taking the square root on both sides completes the proof. □

Chapitre 4

A posteriori energy-norm error estimate based on flux reconstruction

Submitted to SIAM Journal on Numerical Analysis under the title 'Improved energy norm a posteriori error estimation based on flux reconstruction for discontinuous Galerkin methods'.

Alexandre Ern[1], Annette F. Stephansen[1,2] and Martin Vohralík[3]

Abstract: We propose and study a new approach to residual a posteriori error estimation in the discontinuous Galerkin finite element method. The main idea, which consists of constructing an $H(\mathrm{div})$-conforming Raviart–Thomas flux on the basis of the conservative discontinuous Galerkin side fluxes, is first exposed for a pure diffusion second-order elliptic problem. In this case, the classical elementwise residual can be transformed into a higher-order term (sometimes considered separately and called "data oscillation term"), thus fully taking advantage of the spectral degrees of freedom within each element available in the discontinuous Galerkin method. Moreover, the classical estimator based on normal gradient jumps is simultaneously replaced by a comparison of the original and reconstructed diffusive fluxes. Finally, our error bound consists of one last estimator which measures the nonconformity of the actual discrete solution by comparing it to its so-called Oswald interpolate. In the second part of the paper, we extend our results to advection–diffusion–reaction problems, where we introduce an additional convective flux reconstruction. Our

[1]Cermics, Ecole des Ponts, ParisTech, 6 et 8 avenue Blaise Pascal, Champs sur Marne, 77455 Marne la Vallée Cedex 2, France.

[2]Andra, Parc de la Croix-Blanche, 1-7 rue Jean Monnet, 92298 Châtenay-Malabry cedex, France.

[3]Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie (Paris 6), B.C. 187, 4 place Jussieu, 75252 Paris cedex 5, France

estimators are based on an abstract upper bound, which is sharp since it is established for arbitrary conforming reconstructions of the discrete solution itself and of its diffusive and convective fluxes. They yield a guaranteed upper bound since all constants are evaluated, are locally efficient, represent local lower bounds of the classical residual estimators, and numerical examples presented at the end of the paper confirm their accuracy and robustness. Incidentally, the $H(\mathrm{div})$-conforming Raviart–Thomas diffusive and convective flux reconstructions are of independent interest.

4.1 Introduction

Let us consider an advection–diffusion–reaction problem

$$-\nabla\cdot(K\nabla u)+\beta\cdot\nabla u+\mu u = f \quad \text{in } \Omega, \tag{4.1a}$$

$$u = 0 \quad \text{on } \partial\Omega, \tag{4.1b}$$

where $\Omega\subset\mathbb{R}^d$ $(d=2,3)$ is a polygonal (polyhedral) domain, K is a diffusion tensor, β is a velocity field, μ is a reaction function, and f is a source term. Our intention is to study a posteriori energy norm error estimates for the approximation of (4.1a)–(4.1b) by interior-penalty discontinuous Galerkin methods with the twofold objective to derive estimates without undetermined constants and to analyze carefully the robustness of the estimates in several practical important situations, *e.g.*, diffusion heterogeneities, dominant advection, or dominant reaction.

For the pure diffusion problem ((4.1a)–(4.1b) with $\beta=\mu=0$), a posteriori error estimates have now been presented in the literature for all major numerical methods. In particular, for the discontinuous Galerkin (DG) one, residual-based energy norm error estimators can be found in the work of Karakashian and Pascal [62], Becker, Hansbo and Larson [20], and Houston, Schötzau and Wihler [58]. New results, closer in spirit to the present approach since they avoid undetermined constants, appeared in the finalization phase of this paper; they include the works of Ainsworth [5], Kim [63,64], Lazarov, Repin and Tomar [67], Cochez-Dhondt and Nicaise [30], and Ern and Stephansen [49].

Although the residual-based energy norm error estimates in [20,58,62] are proved to be both reliable (yield an upper bound on the difference between the exact and approximate solution) and locally efficient (give local lower bounds for the error as well), there is, in our opinion, still room for improvement. First of all, in all these estimates, various undetermined constants appear. As such, the derived estimators should rather be called error indicators, since they are fully usable for the usual practice of identifying the parts of the computational domain with insufficient precision, but not for the actual control over the

error. Hence, the first motivation for our work was to remedy this inconvenience. Secondly, in all these references, the residual estimator in an element T is given by $c_K h_T \|R(u_h)\|_{0,T}$, where $R(u_h) := f + \nabla\cdot(K\nabla_h u_h)$ is the elementwise residue, h_T is the element diameter, and the constant c_K depends only on K (the modifications of [58] do not influence the basic ideas of what follows). In particular, for piecewise constant K and a DG scheme employing first-order polynomials, this reduces to $c_K h_T \|f\|_{0,T}$. We believe that this is not an optimal estimator. In contrast to this situation, the a posteriori error estimates for mixed finite element or finite volume methods recently derived in [101–103] lead to residual estimators of the form $c_K h_T \|f - \Pi_k(f)\|_{0,T}$, where Π_k is the L^2-orthogonal projection onto piecewise polynomials of degree k ($k = 0$ for finite volumes and it is the scalar unknown polynomial degree for mixed finite elements), which is obviously of one order better for $k = 0$ as soon as f possess an $H^1(T)$ regularity. This result is based on the elementwise conservativity of these methods. Hence, a second motivation for our work was to extend this result to DG methods as well, since these methods are likewise locally conservative. A first result in this direction can be found in [49] where the fact that piecewise constant functions are contained in the DG finite element space is exploited to improve the classical residual estimator to $c_K h_T \|R(u_h) - \Pi_0(R(u_h))\|_{0,T}$, which reduces to $c_K h_T \|f - \Pi_0(f)\|_{0,T}$ if first-order polynomials are used. Finally, it is quite usual in the a posteriori error estimation literature to encounter a residual estimator in each element of the form $c_K h_T \|R(u_h)\|_{0,T}$ and a separate "data oscillation term" $c_K h_T \|f - \Pi_k(f)\|_{0,T}$. In our approach, these two terms are merged into a single residual estimator of the form $c_K h_T \|f - \Pi_k(f)\|_{0,T}$.

One obtains $c_K h_T \|R(u_h)\|_{0,T}$ as the residual term when the elliptic operator is applied directly to the discrete solution u_h, after an integration by parts has been performed. Since the diffusive flux $-K\nabla_h u_h$ of the approximate DG solution u_h is not in $H(\mathrm{div}, \Omega)$, there also appears a so-called mass balance estimator, typically of the form $c_K h_F^{1/2} \|n_F^t [\![K\nabla_h u_h]\!]\|_{0,F}$ for each face F, where h_F is the diameter of F and where $[\![\cdot]\!]$ is the jump operator given by equation (4.2) below. By such a direct approach, one in some sense ignores the local conservativity imbedded in DG schemes. The basic idea of our approach is to first introduce an $H(\mathrm{div}, \Omega)$-conforming reconstruction of the diffusive flux $\mathbf{t}_h$. By suitably choosing $\mathbf{t}_h$ in Raviart–Thomas spaces, used extensively in the mixed finite element method, *cf.* [23, 91], the mass balance estimator is replaced by a comparison of the original and reconstructed diffusive fluxes of the form $\|K^{1/2}\nabla u_h + K^{-1/2}\mathbf{t}_h\|_{0,T}$. We next prove that this new estimator represents a lower bound for the original mass balance estimator plus a part of the classical nonconformity estimator (see below), which, together with the results of the previous paragraph, closes the improvement circle. Lastly, this locally computable estimate is only

one possible realization of the general estimator $\inf_{\mathbf{t}\in H(\mathrm{div},\Omega)} \|K^{1/2}\nabla_h u_h + K^{-1/2}\mathbf{t}\|_{0,\Omega}$ that we show to be optimally efficient.

The last typical DG residual estimator measures the nonconformity in the approximate solution u_h and usually takes the form $c_K h_F^{-1/2}\|[\![u_h]\!]\|_{0,F}$ for each face F. However, it appears unnecessary at the estimation stage to go up to this form. The term $\|K^{1/2}\nabla(u_h - \mathcal{I}_{\mathrm{Os}}(u_h))\|_{0,T}$, with $\mathcal{I}_{\mathrm{Os}}(u_h)$ the Oswald $H_0^1(\Omega)$-conforming interpolate of the original nonconforming u_h, is the usual starting point, it is a lower bound for the above one, and presents the additional advantages that it does not feature any undetermined interpolation constant and that it yields a direct (and correct) dependence on K. Again, the completely general form for this estimator is $\inf_{s\in H_0^1(\Omega)}\|K^{1/2}\nabla_h(u_h - s)\|_{0,\Omega}$.

Estimators based on comparisons with reconstructed $H(\mathrm{div})$-conforming fluxes in the continuous finite element method can be traced back to the ideas of Prager and Synge [81] and include, *e.g.*, the works of Ladevèze [65], Ladevèze and Leguillon [66], Destuynder and Métivet [37] and B. Achchab, S. Achchab, Agouzal and Ellaia [1]. The estimates [5,30,63, 64,67] for DG discretizations of pure diffusion problems develop this way. In particular, Ainsworth [5] gives a fully computable estimate for the symmetric interior-penalty DG scheme in the case $d = 2$, $k = 1$, and $K = Id$ (actually, the reconstructed flux $\mathbf{t}_h$ is not directly computed). Kim in [63] uses an $H(\mathrm{div})$-conforming flux reconstruction and gives an estimate for the original unknown and this reconstruction for $d = 2$. Next, in [64], he presents a result similar to that of Cochez-Dhondt and Nicaise [30] and to the one given here for the pure diffusion case. Finally, Lazarov, Repin and Tomar [67] present essentially numerical experiments for yet a similar estimator.

The setting of the present paper includes a large class of interior-penalty DG schemes. We treat the complete advection–diffusion–reaction case and present an abstract framework, established for arbitrary conforming reconstructions of the discrete solution itself and of its diffusive flux and convective fluxes, and show that this framework is optimal. Our estimates are given in the natural energy semi-norm for the DG approximate solution u_h, which is the energy norm for the flux $-K\nabla_h u_h$. We then prove rigorously the local efficiency of the derived estimators, this time in a norm including a term with jumps. We also pay a special attention to the case of a heterogeneous and anisotropic diffusion tensor K; it turns out that some fully robust results with respect to diffusion heterogeneities can be obtained for our new diffusive flux estimator for a certain class of DG schemes such as those introduced by Ern, Stephansen and Zunino [51]. These schemes use diffusivity-dependent weighted averages to formulate the consistency terms and the harmonic average of normal diffusivity to penalize jumps at interfaces. Next, our error estimates, as well

as the upper and global lower bounds within the abstract framework, do not require the mesh to be shape-regular and the data can be as general functions as possible (the usual requirement of shape-regularity and of polynomial data, or, equivalently, the introduction of higher-order oscillation terms, is only needed for the local efficiency proofs). Also, no saturation assumption, no convexity of Ω, no additional regularity of the weak solution of (4.1a)–(4.1b), and no Helmholtz decomposition are needed in our setting. Finally, we have only considered the homogeneous Dirichlet boundary condition for the sake of simplicity; extensions to heterogeneous Dirichlet and Neumann boundary conditions are possible using the concepts of, *e.g.*, [30, 63, 101]. A similar remark applies also to nonmatching meshes, *cf.*, *e.g.*, [101], while constructing the Oswald interpolate as well as the conforming diffusive and convective flux reconstructions on a matching refinement of the given nonmatching grid.

The paper is organized as follows: we first introduce the schemes, notation, assumptions, and the continuous problems in Section 4.2. We then present the details for the pure diffusion problem. First, in Section 4.3, we state both the abstract (containing the above-discussed infimum over continuous spaces) and locally computable (using particular conforming scalar and diffusive flux reconstructions) forms of our a posteriori error estimates. In Section 4.4, we then show that our abstract framework gives a quasi-optimal global efficiency of $\sqrt{2}$ and that the locally computable estimate is optimal up to heterogeneities and anisotropies. An abstract a posteriori error estimate for the reconstructed diffusive flux itself, which allows to improve the global efficiency to the optimal constant 1, is then given in Section 4.5. In Sections 4.6 and 4.7, we then extend the results of the pure diffusion case to the full advection–diffusion–reaction one. While the abstract upper bound stays quasi-optimal with global efficiency of 2, the presented choice of discrete reconstructions leads only to semi-robust estimates in this case, with local efficiency depending on local variations in the coefficients and on the local Péclet number. Finally, numerical experiments of Section 4.8 confirm the accuracy and robustness of our estimators.

4.2 Notation, assumptions, and continuous and discrete problems

4.2.1 Notation

Let $\{\mathcal{T}_h\}_{h>0}$ be a family of triangulations of the domain Ω, consisting of simplices (triangles if $d = 2$, tetrahedra if $d = 3$). A generic element in $\mathcal{T}_h$ is denoted by T, h_T stands for the

diameter of T, and n_T for its outward unit normal. We suppose that $\mathcal{T}_h$ is matching in the sense that it contains no "hanging nodes", *i.e.*, such that if $T, T' \in \mathcal{T}_h$, $T \neq T'$, then $T \cap T'$ is either an empty set or their common face, edge, or vertex. For the local efficiency proofs of our estimators, we will later need the assumption that $\mathcal{T}_h$ is shape-regular in the sense that there exists a constant $\kappa_{\mathcal{T}} > 0$ such that $\min_{T \in \mathcal{T}_h} |T|/h_T^d \geq \kappa_{\mathcal{T}}$ for all $h > 0$. We will be using the "broken Sobolev space" $H^s(\mathcal{T}_h)$,

$$H^s(\mathcal{T}_h) := \{v \in L^2(\Omega); v|_T \in H^s(T) \quad \forall T \in \mathcal{T}_h\},$$

along with its DG approximation space

$$V_h^k := \{v_h \in L^2(\Omega); v_h|_T \in \mathbb{P}_k(T) \quad \forall T \in \mathcal{T}_h\},$$

where $\mathbb{P}_k(T)$ is the set of polynomials of degree less than or equal to k on an element T, $k \geq 1$.[4] The L^2-scalar product and its associated norm on a region $R \subset \Omega$ are indicated by the subscript $0, R$; shall R coincide with Ω, this subscript will be dropped off. For $s \geq 1$, a norm (semi-norm) with the subscript s, R designates the usual norm (semi-norm) in $H^s(R)$. Finally, we use the symbol $\nabla_h v_h$ in order to denote the piecewise gradient of $v \in H^1(\mathcal{T}_h)$, that is, $\nabla_h v \in [L^2(\Omega)]^d$ and for all $T \in \mathcal{T}_h$, $(\nabla_h v)|_T = \nabla(v|_T)$.

We say that F is an interior face of the mesh if there are $T^-(F)$ and $T^+(F)$ in $\mathcal{T}_h$ such that $F = T^-(F) \cap T^+(F)$ and we let n_F be the unit normal vector to F pointing from $T^-(F)$ towards $T^+(F)$. Similarly, we say that F is a boundary face of the mesh if there is $T(F) \in \mathcal{T}_h$ such that $F = T(F) \cap \partial\Omega$ and we let n_F coincide with the outward normal to $\partial\Omega$. All the interior (resp., boundary) faces of the mesh are collected into the set $\mathcal{F}_h^i$ (resp., $\mathcal{F}_h^{\partial\Omega}$) and we let $\mathcal{F}_h = \mathcal{F}_h^i \cup \mathcal{F}_h^{\partial\Omega}$; $\mathcal{F}_T$ is then the set of faces of a given $T \in \mathcal{T}_h$ and $\widetilde{\mathcal{F}}_T$ is the set of such faces that share at least a vertex with T. Similarly, $\mathcal{T}_T$ is the set including the simplex T and its neighbors and $\widetilde{\mathcal{T}}_T$ contains all $T' \in \mathcal{T}_h$ that share at least a vertex with T (including T itself). Henceforth, we shall often deal with functions that are double-valued on $\mathcal{F}_h^i$ and single-valued on $\mathcal{F}_h^{\partial\Omega}$. This is the case, for instance, of functions in V_h^k. On interior faces, when the two branches of the function in question, say v, are associated with restrictions to the neighboring elements $T^{\mp}(F)$, these branches are denoted by $v^{\mp}$ and the jump of v across F is defined as

$$[\![v]\!]_F := v^- - v^+. \tag{4.2}$$

On an interior face $F \in \mathcal{F}_h^i$, we define the standard (arithmetic) average as $\{v\}_F := \frac{1}{2}(v^- + v^+)$; the subscript F in the above jumps and averages is omitted if there is no

[4] Dans les chapitres précédents, k était noté p.

ambiguity. For convenience, we set $[\![v]\!]_F := v|_F$ and $\{v\}_F := \frac{1}{2}v|_F$ on boundary faces. Finally, the weighted average of a two-valued function on an interior face $F \in \mathcal{F}_h^i$ is defined as

$$\{v\}_\omega := \omega_{T^-(F),F} v^- + \omega_{T^+(F),F} v^+, \tag{4.3}$$

where the nonnegative weights have to satisfy $\omega_{T^-(F),F} + \omega_{T^+(F),F} = 1$. On boundary faces, we set $\{v\}_\omega := v$ and $\omega_{T(F),F} := 1$. Finally, for all $T \in \mathcal{T}_h$ and $F \in \mathcal{F}_T$, we let $\bar{\omega}_{T,F} := 1 - \omega_{T,F}$.

4.2.2 Assumptions

We suppose in this paper that $K \in [L^\infty(\Omega)]^{d\times d}$ is a symmetric, uniformly positive definite, and piecewise constant tensor and we denote by $\lambda_{m,T}$ and $\lambda_{M,T}$, respectively, its minimum and maximum eigenvalue on $T \in \mathcal{T}_h$. Next, $\beta \in H(\mathrm{div},\Omega) \cap [L^\infty(\Omega)]^d$, $\mu \in L^\infty(\Omega)$, and $\mu - \frac{1}{2}\nabla\cdot\beta \geq 0$ are supposed and we use $\tilde{\mu}_{m,T}$ to indicate the (essential) minimum value of $\mu - \frac{1}{2}\nabla\cdot\beta$ on T; we also suppose that if $\tilde{\mu}_{m,T} = 0$, then $\|\mu\|_{\infty,T} = \|\frac{1}{2}\nabla\cdot\beta\|_{\infty,T} = 0$. Finally, $f \in L^2(\Omega)$ is supposed. These assumptions will be sufficient for the existence and uniqueness of both continuous and discrete problems and for our a posteriori error estimates, as well as for the global efficiency of the abstract estimates; for the present proof of the local efficiency of the locally computable estimates, however, we shall later tighten them.

4.2.3 The continuous problem

We define the bilinear form B by

$$B(u,v) := \sum_{T\in\mathcal{T}_h} \left\{ (K\nabla u, \nabla v)_{0,T} + (\beta\cdot\nabla u, v)_{0,T} + (\mu u, v)_{0,T} \right\} \qquad u, v \in H^1(\mathcal{T}_h) \tag{4.4}$$

and the corresponding energy (semi-)norm by

$$\|v\|_B^2 := \sum_{T\in\mathcal{T}_h} \|v\|_{B,T}^2, \quad \|v\|_{B,T}^2 := \|K^{\frac{1}{2}}\nabla v\|_{0,T}^2 + \left\| \left(\mu - \tfrac{1}{2}\nabla\cdot\beta\right)^{\frac{1}{2}} v \right\|_{0,T}^2 \qquad v \in H^1(\mathcal{T}_h). \tag{4.5}$$

We remark that $\|\cdot\|_B$ is always a norm on $H_0^1(\Omega)$, whereas it is a norm on $H^1(\mathcal{T}_h)$ only when $\tilde{\mu}_{m,T} > 0$ for all $T \in \mathcal{T}_h$.

The weak formulation of the problem (4.1a)–(4.1b) is then to find $u \in H_0^1(\Omega)$ such that

$$B(u,v) = (f,v) \qquad \forall v \in H_0^1(\Omega). \tag{4.6}$$

The assumptions of the previous section, the Green theorem, and the Cauchy–Schwarz inequality imply that

$$\begin{aligned} B(v,v) &= \|v\|_B^2 \qquad \forall v \in H_0^1(\Omega), && (4.7)\\ B(u,v) &\le \max\left\{1, \max_{T\in\mathcal{T}_h}\left\{\frac{\|\mu\|_{\infty,T}}{\tilde{\mu}_{m,T}}\right\}\right\}\|u\|_B\|v\|_B \\ &\quad + \max_{T\in\mathcal{T}_h}\left\{\frac{\|\beta\|_{\infty,T}}{\lambda_{m,T}^{1/2}}\right\}\|u\|_B\|v\|_{0,\Omega} \qquad \forall u,v \in H^1(\mathcal{T}_h). && (4.8)\end{aligned}$$

Hence, problem (4.6) admits a unique solution.

Remark (Notation). If $\tilde{\mu}_{m,T} = 0$, then the term $\|\mu\|_{\infty,T}/\tilde{\mu}_{m,T}$ in estimate (4.8) should be evaluated as zero, since in this case we assume $\|\mu\|_{\infty,T} = 0$. To simplify the notation, we will systematically use the convention $0/0 = 0$ throughout the text.

4.2.4 The discontinuous Galerkin method

The interior-penalty DG methods considered in this paper are associated with the bilinear form

$$\begin{aligned} B_h(u,v) &:= (K\nabla_h u, \nabla_h v) + ((\mu - \nabla\cdot\beta)u, v) - (u, \beta\cdot\nabla_h v) \\ &\quad - \sum_{F\in\mathcal{F}_h}\left\{(n_F^t\{K\nabla_h u\}_\omega, [\![v]\!])_{0,F} + \theta(n_F^t\{K\nabla_h v\}_\omega, [\![u]\!])_{0,F}\right\} \qquad (4.9)\\ &\quad + \sum_{F\in\mathcal{F}_h}\left\{(\gamma_F[\![u]\!], [\![v]\!])_{0,F} + (\beta\cdot n_F\{u\}, [\![v]\!])_{0,F}\right\}. \end{aligned}$$

The discrete problem now consists of finding $u_h \in V_h^k$ such that

$$B_h(u_h, v_h) = (f, v_h) \qquad \forall v_h \in V_h^k. \tag{4.10}$$

Taking in (4.9) the weights on interior faces equal to $1/2$ and letting $\theta = 0$, $\theta = -1$, or $\theta = 1$ leads to the well-known Incomplete, Nonsymmetric, or Symmetric Interior-Penalty discontinuous Galerkin methods. The stabilization parameter γ_F takes the general form

$$\gamma_F := \alpha_F\frac{\gamma_{K,F}}{h_F} + \gamma_{\beta,F} \qquad \forall F \in \mathcal{F}_h, \tag{4.11}$$

where α_F is a (user-dependent) positive parameter, $\gamma_{K,F}$ a positive scalar-valued function depending on K, and $\gamma_{\beta,F}$ a nonnegative scalar-valued function depending on β and vanishing if $\beta = 0$ (the usual choice is $\gamma_{\beta,F} = \frac{1}{2}|\beta\cdot n_F|$, which amounts to so-called upwinding).

As usual with interior-penalty methods, the parameters α_F must be taken large enough to ensure the coercivity of the discrete bilinear form B_h on V_h^k whenever $\theta \neq -1$. Some additional assumptions on the weights and the penalty coefficient $\gamma_{K,F}$ will be introduced later in order to ensure the robustness of our estimates with respect to diffusion heterogeneities; see Theorems 4.7 and 4.18 below. The recently derived weighted interior penalty DG method of [51] satisfies these assumptions.

4.3 Improved energy norm a posteriori error estimates in the pure diffusion case

We present in this section our a posteriori estimates on the error between the weak solution u and the DG approximate solution u_h in the pure diffusion case. Note that at this stage, neither additional assumptions on the data (in particular, f need not be a polynomial) nor the shape-regularity of the mesh are required.

4.3.1 Abstract framework

The following lemma gives the basic abstract framework for our a posteriori error estimates in the pure diffusion case. It follows from [63, Lemma 4.4] and it is analogous to, but simpler than, the Helmholtz decomposition of, *e.g.*, [5, Theorem 1]; a similar but more general result, applicable also in the advection–diffusion–reaction case (and used in Section 4.6 below) is given in [102, Lemma 7.1].

Lemma 4.1 (Abstract framework in the pure diffusion case)**.** *Let $\beta = \mu = 0$ and let $u \in H_0^1(\Omega)$ and $u_h \in H^1(\mathcal{T}_h)$ be arbitrary. Then*

$$\|u - u_h\|_B^2 \leq \inf_{s \in H_0^1(\Omega)} \|u_h - s\|_B^2 + \sup_{\varphi \in H_0^1(\Omega),\, \|\varphi\|_B = 1} B(u - u_h, \varphi)^2. \tag{4.12}$$

Proof. Following [63, Lemma 4.4], let $\psi \in H_0^1(\Omega)$ be such that

$$B(\psi, v) = B(u_h, v) \qquad \forall v \in H_0^1(\Omega).$$

Then

$$\|u - u_h\|_B^2 = \|u_h - \psi\|_B^2 + B\left(u - u_h, \frac{u - \psi}{\|u - \psi\|_B}\right)^2,$$

whence the conclusion is straightforward. □

4.3.2 Abstract a posteriori error estimate

We next give here an abstract form of our a posteriori error estimate.

Theorem 4.2 (Abstract a posteriori error estimate in the pure diffusion case)**.** *Let $\beta = \mu = 0$, let u be the unique solution of (4.6), and let $u_h \in H^1(\mathcal{T}_h)$ be arbitrary. Then*

$$\begin{aligned} \|u - u_h\|_B^2 \quad &\leq \quad \inf_{s \in H_0^1(\Omega)} \|u_h - s\|_B^2 \\ &\quad + \inf_{\mathbf{t} \in H(\mathrm{div},\Omega)} \sup_{\varphi \in H_0^1(\Omega),\, \|\varphi\|_B = 1} ((f - \nabla{\cdot}\mathbf{t}, \varphi) - (K\nabla_h u_h + \mathbf{t}, \nabla\varphi))^2. \end{aligned} \tag{4.13}$$

Proof. By (4.6), we immediately have $B(u, \varphi) = (f, \varphi)$. Using this we obtain, for an arbitrary $\mathbf{t} \in H(\mathrm{div}, \Omega)$ and employing the Green theorem,

$$\begin{aligned} B(u - u_h, \varphi) \quad &= \quad (f, \varphi) - (K\nabla_h u_h, \nabla\varphi) = (f, \varphi) - (K\nabla_h u_h + \mathbf{t}, \nabla\varphi) + (\mathbf{t}, \nabla\varphi) \\ &= \quad (f - \nabla{\cdot}\mathbf{t}, \varphi) - (K\nabla_h u_h + \mathbf{t}, \nabla\varphi). \end{aligned}$$

□

Remark (Form of the abstract estimate of Theorem 4.2). It has been already noted in, *e.g.*, [2,3,41,63,102] that the first term of (4.12) evaluates the "nonconforming error" in the scalar unknown u_h. The second term of (4.12) is called in [3,63] the "conforming error". Relation (4.13) actually shows that this second term is related to the residual and to the nonconformity in the flux $-K\nabla_h u_h$.

Remark (A first computable a posteriori error estimate). We remark that using the Cauchy–Schwarz inequality, the Friedrichs inequality $\|\varphi\|^2 \leq C_{\mathrm{F},\Omega} h_\Omega^2 \|\nabla\varphi\|^2$, the definition (4.5) of the energy semi-norm, and the fact that $\|\varphi\|_B = 1$, it follows readily from (4.13) that

$$\|u - u_h\|_B^2 \leq \|u_h - s\|_B^2 + \left(\frac{C_{\mathrm{F},\Omega}^{1/2} h_\Omega}{\min_{T \in \mathcal{T}_h} \lambda_{m,T}^{1/2}} \|f - \nabla{\cdot}\mathbf{t}\|_{0,\Omega} + \|K^{\frac{1}{2}} \nabla_h u_h + K^{-\frac{1}{2}} \mathbf{t}_h\|_{0,\Omega} \right)^2$$

for any $s \in H_0^1(\Omega)$ and any $\mathbf{t} \in H(\mathrm{div}, \Omega)$. This is an estimate similar to that proposed by Lazarov, Repin and Tomar [67]. As promoted in [67], this estimate is scheme-independent. On the other hand, being scheme-independent means that we are not using all the information that we have once the computation has been finished. As we will see later, this information can be used to improve the residual. Another disadvantage of the above estimate is that the dependence on the diffusion tensor K is very unfavorable in the presence of strong heterogeneities; this point was however not addressed in [67].

Although Theorem 4.2 gives a framework for a quasi-optimal a posteriori error estimate (see Section 4.4.1 below), such an estimate is not practically computable. To this purpose, we have to choose a particular $s \in H_0^1(\Omega)$ and $\mathbf{t} \in H(\mathrm{div}, \Omega)$. We devote the two following sections to this point.

4.3.3 Oswald interpolation operator

The Oswald interpolate of u_h was already used as a suitable $s \in H_0^1(\Omega)$ in a posteriori error estimation in nonconforming or DG methods, *cf.* [2, 41, 62]. It has been analyzed in detail in [24, 62]. The Oswald interpolation operator $\mathcal{I}_{\mathrm{Os}} : V_h^k \to V_h^k \cap H_0^1(\Omega)$ is defined as follows: given a function $v_h \in V_h^k$, the value of $\mathcal{I}_{\mathrm{Os}}(v_h)$ is prescribed at suitable (*e.g.*, Lagrangian) vertices of the simplices of $\mathcal{T}_h$. At the vertices located inside Ω, the average of the values of v_h is specified by

$$\mathcal{I}_{\mathrm{Os}}(v_h)(V) = \frac{1}{|\mathcal{T}_V|} \sum_{T \in \mathcal{T}_V} v_h|_T(V),$$

where $\mathcal{T}_V$ is the set of $T \in \mathcal{T}_h$ that contain the vertex V, while at boundary vertices, the value of $\mathcal{I}_{\mathrm{Os}}(v_h)$ is set to zero. The following results have been proved in [24, Lemma 3.2 and Remark 3.2] and [62, Theorem 2.2]:

Lemma 4.3 (Oswald interpolation operator)**.** *Let $\mathcal{T}_h$ be shape-regular, let $v_h \in V_h^k$, and let $\mathcal{I}_{\mathrm{Os}}(v_h)$ be constructed as above. Then,*

$$\begin{aligned} \|v_h - \mathcal{I}_{\mathrm{Os}}(v_h)\|_{0,T}^2 &\leq C \sum_{F \in \widetilde{\mathcal{F}}_T} h_F \|[\![v_h]\!]\|_{0,F}^2, \\ \|\nabla(v_h - \mathcal{I}_{\mathrm{Os}}(v_h))\|_{0,T}^2 &\leq C \sum_{F \in \widetilde{\mathcal{F}}_T} h_F^{-1} \|[\![v_h]\!]\|_{0,F}^2, \end{aligned}$$

where the constants C only depend on the space dimension d, the maximal polynomial degree k, and the shape regularity parameter $\kappa_{\mathcal{T}}$.

4.3.4 Diffusive flux reconstruction

A choice of suitable $\mathbf{t} \in H(\mathrm{div}, \Omega)$ in Theorem 4.2 is a more delicate question. Remark in particular that $\mathbf{t} \in H(\mathrm{div}, \Omega)$ is a necessary condition but some result on the divergence of $\mathbf{t}$ will also be necessary in view of the abstract a posteriori error estimate. A previous work on $H(\mathrm{div}, \Omega)$ flux postprocessing in DG methods includes the paper of Bastian and Rivière [18], but we shall choose here the postprocessing recently derived in [46] or in [64].

To this purpose, we will need the Raviart–Thomas–Nédélec spaces of vector functions (*cf.* [23, 74, 82, 91])

$$\begin{aligned}\mathbf{RTN}_T^l &= \mathbb{P}_l^d(T) + x\mathbb{P}_l(T),\\ \mathbf{RTN}_h^l &= \big\{v_h \in H(\mathrm{div},\Omega)\,;\, v_h|_T \in \mathbf{RTN}_T^l \quad \forall T\in\mathcal{T}_h\big\}.\end{aligned}$$

In particular, $v_h \in \mathbf{RTN}_h^l$ is such that $\nabla{\cdot}v_h \in \mathbb{P}_l(T)$ for all $T\in\mathcal{T}_h$, $v_h{\cdot}n_F \in \mathbb{P}_l(F)$ for all $F\in\mathcal{F}_T$ and all $T\in\mathcal{T}_h$, and such that its normal trace is continuous.

Using the specification of the degrees of freedom of functions in $\mathbf{RTN}_T^l$ given in the above citations, our $H(\mathrm{div},\Omega)$-conforming diffusive flux reconstruction $\mathbf{t}_h$ will belong to $\mathbf{RTN}_h^l$ with $l=k$ or $l=k-1$ and we prescribe it locally on all $T\in\mathcal{T}_h$ as follows:

$$(\mathbf{t}_h{\cdot}n_F, q_h)_{0,F} = \left(-n_F^t\{K\nabla_h u_h\}_\omega + \alpha_F\frac{\gamma_{K,F}}{h_F}[\![u_h]\!], q_h\right)_{0,F} \qquad \forall q_h\in\mathbb{P}_l(F),\ \forall F\in\mathcal{F}_T, \tag{4.14}$$

$$(\mathbf{t}_h, r_h)_{0,T} = -(K\nabla u_h, r_h)_{0,T} + \theta\sum_{F\in\mathcal{F}_T}\omega_{T,F}(n_F^tKr_h, [\![u_h]\!])_{0,F} \quad \forall r_h\in\mathbb{P}_{l-1}^d(T). \tag{4.15}$$

Note in particular that the quantities prescribing the moments of $\mathbf{t}_h{\cdot}n_F$ are uniquely defined for each face $F\in\mathcal{F}_h$, whence the continuity of the normal trace of $\mathbf{t}_h$. By this construction, we have the following crucial lemma:

Lemma 4.4 (Reconstructed diffusion residual)**.** *There holds*

$$(\nabla{\cdot}\mathbf{t}_h, \xi_h)_{0,T} = (f,\xi_h)_{0,T} \qquad \forall T\in\mathcal{T}_h,\ \forall\xi_h\in\mathbb{P}_l(T), \tag{4.16}$$

which yields, using that $\nabla{\cdot}\mathbf{t}_h|_T\in\mathbb{P}_l(T)$*,*

$$\nabla{\cdot}\mathbf{t}_h|_T = \Pi_l(f)|_T \qquad \forall T\in\mathcal{T}_h.$$

Proof. Let $\xi_h\in\mathbb{P}_l(T)$ be arbitrary. We then have, using the Green theorem, the fact that $\xi_h|_F\in\mathbb{P}_l(F)$ for all $F\in\mathcal{F}_T$, $\nabla\xi_h\in\mathbb{P}_{l-1}^d(T)$, the definition (4.14)–(4.15) of $\mathbf{t}_h$, and putting $v_h=\xi_h$ on T and $v_h=0$ otherwise,

$$\begin{aligned}(\nabla{\cdot}\mathbf{t}_h,\xi_h)_{0,T} &= -(\mathbf{t}_h,\nabla\xi_h)_{0,T} + \sum_{F\in\mathcal{F}_T}(\mathbf{t}_h{\cdot}n_T,\xi_h)_{0,F} = (K\nabla_h u_h,\nabla_h v_h)_{0,T}\\ &\quad - \sum_{F\in\mathcal{F}_T}\Big\{\theta(n_F^t\{K\nabla_h v_h\}_\omega, [\![u_h]\!])_{0,F}\\ &\qquad + \Big(n_F^t\{K\nabla_h u_h\}_\omega - \alpha_F\frac{\gamma_{K,F}}{h_F}[\![u_h]\!], [\![v_h]\!]\Big)_{0,F}\Big\}\\ &= B_h(u_h,v_h) = (f,v_h) = (f,\xi_h)_{0,T},\end{aligned}$$

employing finally the definition of the DG bilinear form (4.9) and that of the DG approximate solution (4.10). □

4.3.5 Locally computable a posteriori error estimate

With the results of the three previous sections, we are now ready to state our practical locally computable a posteriori error estimate for DG methods.

Let us define the *nonconformity estimator* $\eta_{\mathrm{NC},T}$ by

$$\eta_{\mathrm{NC},T} := \|u_h - \mathcal{I}_{\mathrm{Os}}(u_h)\|_{B,T}, \tag{4.17}$$

and the *diffusive flux estimator* $\eta_{\mathrm{DF},T}$ by

$$\eta_{\mathrm{DF},T} \quad := \quad \|K^{\frac{1}{2}}\nabla u_h + K^{-\frac{1}{2}}\mathbf{t}_h\|_{0,T}, \tag{4.18}$$

where $\mathbf{t}_h \in \mathbf{RTN}_h^l$ is given by (4.14)–(4.15). Finally, let us put

$$m_{T,K}^2 := C_{\mathrm{P}}\frac{h_T^2}{\lambda_{m,T}}$$

for all $T \in \mathcal{T}_h$, where C_{P} is the constant from the Poincaré inequality

$$\|\varphi - \Pi_0(\varphi)\|_{0,T}^2 \le C_{\mathrm{P}} h_T^2 \|\nabla\varphi\|_{0,T}^2 \qquad \forall\varphi \in H^1(T), \tag{4.19}$$

which can be evaluated as $1/\pi^2$ owing to the convexity of $T \in \mathcal{T}_h$, *cf.* [19, 79]. We define the *residual estimator* $\eta_{\mathrm{R},T}$ by

$$\eta_{\mathrm{R},T} := m_{T,K}\|f - \Pi_l(f)\|_{0,T}. \tag{4.20}$$

We then have the following a posteriori error estimate:

Theorem 4.5 (Locally computable a posteriori error estimate in the pure diffusion case)**.** *Let $\beta = \mu = 0$, let u be the unique solution of* (4.6)*, and let u_h be its discontinuous Galerkin approximation given by* (4.10)*. Then*

$$\|u - u_h\|_B \le \left\{\sum_{T\in\mathcal{T}_h}\left\{\eta_{\mathrm{NC},T}^2 + (\eta_{\mathrm{R},T} + \eta_{\mathrm{DF},T})^2\right\}\right\}^{1/2}.$$

Proof. Put $s = \mathcal{I}_{\mathrm{Os}}(u_h)$ and $\mathbf{t} = \mathbf{t}_h$ in Theorem 4.2. Note that, for each $T \in \mathcal{T}_h$,

$$\begin{aligned} |(f - \nabla\cdot\mathbf{t}_h, \varphi)_{0,T}| &= |(f - \nabla\cdot\mathbf{t}_h, \varphi - \Pi_0(\varphi))_{0,T}| \\ &= |(f - \Pi_l(f), \varphi - \Pi_0(\varphi))_{0,T}| \le \eta_{\mathrm{R},T}\|\varphi\|_{B,T}, \end{aligned} \tag{4.21}$$

using Lemma 4.4, the Poincaré inequality 4.19, and the definition (4.5) of the energy norm (note that this step holds true for both $l = k$ or $l = k-1$). Next, $|(K\nabla_h u_h + \mathbf{t}, \nabla\varphi)_{0,T}| \leq \eta_{\mathrm{DF},T}\|\varphi\|_{B,T}$ is immediate. Hence it now suffices to use the Cauchy–Schwarz inequality and to notice that $\|\varphi\|_B = 1$ in order to conclude the proof. □

Remark (Properties of the estimate of Theorem 4.5). The following properties of the estimate of Theorem 4.5 can be mentioned:

- It gives a guaranteed upper bound, *i.e.*, features no undetermined constant.
- The residual estimator $\eta_{\mathrm{R},T}$ coincides with the classical (properly weighted) "data oscillation term", whence it represents a major improvement of the classical residual estimator, which is of the form $c_K h_T\|f + \nabla\cdot(K\nabla_h u_h)\|_{0,T}$. Also, although it represents a higher-order term for piecewise smooth f, it shall not be neglected as it can be important on coarse grids or for highly varying K.
- The Poincaré constant C_{P} does not depend on the shape-regularity of the mesh, whence the present estimate is valid also on anisotropic meshes.
- The Poincaré constant C_{P} does not depend on the polynomial degree of u_h, so that, in contrast to the estimates of [20, 62], the present estimate is valid uniformly with respect to k.
- No assumption on the polynomial form of f is needed at this stage.
- Letting $\eta_{\mathrm{R},T} = m_{T,K}\|f - \nabla\cdot\mathbf{t}_h\|_{0,T}$, the present estimate is valid for any $\mathbf{t}_h \in H(\mathrm{div},\Omega)$ such that $(\nabla\cdot\mathbf{t}_h, 1)_{0,T} = (f,1)_{0,T}$ for all $T \in \mathcal{T}_h$, which is a local (conservativity) property, in contrast to the global Galerkin orthogonality used traditionally for conforming finite element methods.

4.4 Efficiency of the estimates in the pure diffusion case

In this section, we first rapidly check the (global) efficiency of the abstract framework of Theorem 4.2. We then investigate in detail the (local) efficiency of the a posteriori error estimate of Theorem 4.5.

4.4.1 Global efficiency of the abstract estimate

Theorem 4.6 (Global efficiency of the abstract estimate in the pure diffusion case)**.** *Let $\beta = \mu = 0$, let u be the unique solution of (4.6), and let $u_h \in H^1(\mathcal{T}_h)$ be arbitrary. Let the a posteriori error estimate be given by Theorem 4.2. Then*

$$\inf_{s\in H_0^1(\Omega)} \|u_h - s\|_B^2 + \inf_{\mathbf{t}\in H(\mathrm{div},\Omega)} \sup_{\varphi\in H_0^1(\Omega),\, \|\varphi\|_B=1} ((f - \nabla\cdot\mathbf{t}, \varphi) - (K\nabla_h u_h + \mathbf{t}, \nabla\varphi))^2 \leq 2\|u - u_h\|_B^2.$$

Proof. It suffices to put $s = u$, $\mathbf{t} = -K\nabla u$, and to use the Cauchy–Schwarz inequality and the fact that $\|\varphi\|_B = 1$. □

Remark (Global effectivity index). It follows from Theorem 4.6 that the abstract a posteriori error estimate of Theorem 4.2 is quasi-exact in the sense that the effectivity index, *i.e.*, the ratio of the estimated to the actual error, is equal to $\sqrt{2}$. The effectivity index can be improved to 1 when either $u_h \in H_0^1(\Omega)$ or $-K\nabla_h u_h \in H(\mathrm{div},\Omega)$, but this is not to be expected apart from particular cases. A possible remedy is presented in Section 4.5 below.

Remark (Robustness with respect to data, polynomial degree, and meshes). It follows from Theorem 4.6 that the abstract a posteriori error estimate of Theorem 4.2 is fully robust with respect to K without any assumption on its distribution, with respect to f (no polynomial form needed), with respect to the polynomial degree k, and finally with respect to the mesh (which can be anisotropic), in the sense that the effectivity index does not depend on these quantities.

4.4.2 Local efficiency of the locally computable estimate

We now investigate how the quasi-optimal abstract global efficiency of the previous section persists for our particular choices of the conforming reconstructions of the discrete solution and of its diffusive flux. To this purpose, we restrict the class of interior-penalty DG schemes by the following assumptions: there exist constants C_1, C_2, and C_3, independent

of K, such that

$$C_1 \min(\lambda_{m,T^+(F)}, \lambda_{m,T^-(F)}) \le \gamma_{K,F} \le C_2 \min(\lambda_{M,T^+(F)}, \lambda_{M,T^-(F)}) \qquad \forall F \in \mathcal{F}_h^i, \tag{4.22}$$

$$C_1 \, \lambda_{m,T(F)} \le \gamma_{K,F} \le C_2 \, \lambda_{M,T(F)} \qquad \forall F \in \mathcal{F}_h^{\partial\Omega}, \tag{4.23}$$

$$n_F^t K n_F \omega_{T(F),F} \le C_3 \gamma_{K,F} \qquad \forall T \in T(F), F \in \mathcal{F}_h^i. \tag{4.24}$$

An example of a DG scheme satisfying (4.22)–(4.24) with $C_1 = 1/2$, $C_2 = 1$, and $C_3 = 1$ is that recently derived by Ern, Stephansen and Zunino [51]. It consists of setting

$$\gamma_{K,F} := \frac{\delta_{K,F+}\delta_{K,F-}}{\delta_{K,F+} + \delta_{K,F-}} \qquad \forall F \in \mathcal{F}_h^i, \tag{4.25}$$

$$\gamma_{K,F} := \delta_{K,F} \qquad \forall F \in \mathcal{F}_h^{\partial\Omega}, \tag{4.26}$$

where $\delta_{K,F\mp} = n_F^t K|_{T^\mp(F)} n_F$ if $F \in \mathcal{F}_h^i$ and $\delta_{K,F} = n_F^t K|_{T(F)} n_F$ if $F \in \mathcal{F}_h^{\partial\Omega}$, while the weights are chosen so that

$$\omega_{T^-(F),F} := \frac{\delta_{K,F+}}{\delta_{K,F+} + \delta_{K,F-}}, \qquad \omega_{T^+(F),F} := \frac{\delta_{K,F-}}{\delta_{K,F+} + \delta_{K,F-}} \qquad \forall F \in \mathcal{F}_h^i. \tag{4.27}$$

Theorem 4.7 (Local efficiency of the locally computable estimate in the pure diffusion case)**.** *Let $\beta = \mu = 0$, let $\mathcal{T}_h$ be shape-regular, let f be a piecewise polynomial of degree m, let u be the unique solution of* (4.6)*, and let u_h be its discontinuous Galerkin approximation given by* (4.10) *with the weights $\omega_{T,F}$ and penalty parameters $\gamma_{K,F}$ (for simplicity supposed facewise constant) satisfying* (4.22)–(4.24)*. Let next the a posteriori error estimate be given by Theorem 4.5, with in particular $\eta_{\mathrm{NC},T}$ given by* (4.17) *and $\eta_{\mathrm{DF},T}$ given by* (4.18)*. Let us put*

$$\lambda_{m,\widetilde{\mathcal{T}}_T} := \min_{T' \in \widetilde{\mathcal{T}}_T} \lambda_{m,T'}, \tag{4.28}$$

and

$$\|v\|_{B,*,\mathcal{F}}^2 := \sum_{F \in \mathcal{F}} \|\gamma_F^{1/2} [\![v]\!]\|_{0,F}^2 \qquad v \in H^1(\mathcal{T}_h), \tag{4.29}$$

where we will either take $\mathcal{F} = \mathcal{F}_h$, $\mathcal{F} = \mathcal{F}_T$, or $\mathcal{F} = \widetilde{\mathcal{F}}_T$. Then, for each $T \in \mathcal{T}_h$, there

holds

$$\eta_{\mathrm{NC},T} \leq C \frac{\lambda_{M,T}^{1/2}}{\lambda_{m,\widetilde{\mathcal{T}}_T}^{1/2}} \|u - u_h\|_{B,*,\widetilde{\mathcal{F}}_T}, \tag{4.30}$$

$$\eta_{\mathrm{DF},T} \leq \tilde{C} \max_{T' \in \mathcal{T}_T} \left\{ \frac{\lambda_{M,T'}}{\lambda_{m,T'}} \right\} \|u - u_h\|_{B,\mathcal{T}_T} + \tilde{C} \left(\frac{\lambda_{M,T}}{\lambda_{m,T}} \right)^{\frac{1}{2}} \|u - u_h\|_{B,*,\mathcal{F}_T}, \tag{4.31}$$

where the constant C depends only on the space dimension d, on the maximal polynomial degree k, on the DG parameters α_F, on the shape regularity parameter $\kappa_{\mathcal{T}}$, and on the constant C_1 from (4.22)–(4.23) *and $\tilde{C}$ in addition depends on the polynomial degree m of f, on the DG parameter θ, and on the constants C_2–C_3 from* (4.22)–(4.24).

Proof. Combine Lemmas 4.8 and 4.10 given below. □

Remark (Local efficiency norm and global efficiency with respect to the energy semi-norm). The local efficiency stated in Theorem 4.7 is given for the energy semi-norm augmented by the natural DG jump semi-norm $\|\cdot\|_{B,*,\mathcal{F}_h}$. Owing to the result of Ainsworth [5, Theorem 3], global efficiency of our nonconformity and diffusive flux estimators in the energy semi-norm $\|\cdot\|_B$ follows from (4.30)–(4.31) for sufficiently large stabilization parameters α_F in the case $d = 2$, $k = 1$, $K = Id$, and $\theta = 1$.

Remark (Efficiency of the residual estimator). We recall that the residual estimator $\eta_{\mathrm{R},T}$ coincides with the usual "data oscillation term" and is in general of higher order, whence no efficiency is to be shown.

Remark (Robustness with respect to heterogeneities and anisotropies). Owing to (4.31), our diffusive flux estimator $\eta_{\mathrm{DF},T}$ is fully robust with respect to diffusion heterogeneities. This is an important property in practical applications, *e.g.*, when dealing with underground flows. The design conditions (4.22)–(4.24) play a crucial role in this respect, *cf.* the proof of Lemma 4.10 below. A similar result was proved recently in [49] in the context of residual-based a posteriori error estimates for DG methods with diffusivity-dependent weights and penalty parameter based on the harmonic average of the normal diffusivity, see (4.25)–(4.27). We next point out that under the assumption of "monotonicity around vertices" distribution of the heterogeneities and using the concepts of, *e.g.*, Ainsworth [3], also the nonconformity estimator $\eta_{\mathrm{NC},T}$ may be shown robust with respect to heterogeneities. Finally, no robustness with respect to anisotropies is achieved by our estimators $\eta_{\mathrm{DF},T}$ and $\eta_{\mathrm{NC},T}$, but, at least, the local efficiency estimates only depend on local, elementwise, anisotropies.

Remark (Generalization to other DG schemes). Making appropriate changes in the proof of Theorem 4.7 below, all the presented results (up to the robustness with respect to heterogeneities) extend appropriately to all the DG schemes included in the setting (4.9)–(4.10), even if the design conditions (4.22)–(4.24) are not satisfied.

Remark (Lower bound for the classical estimator). The proof of Lemma 4.8 below shows that the nonconformity estimator $\eta_{\mathrm{NC},T}$ represents a lower bound for the classical nonconformity estimator $\big\{\sum_{F\in\widetilde{\mathcal{F}}_T} h_F^{-1}\|[\![u_h]\!]\|_{0,F}^2\big\}^{1/2}$. Similarly, the proof of Lemma 4.10 below shows that the diffusive flux estimator $\eta_{\mathrm{DF},T}$ represents a lower bound for the classical gradient jump estimator $\big\{\sum_{F\in\mathcal{F}_T} \|n_F^t[\![K\nabla_h u_h]\!]\|_{0,F}^2\big\}^{1/2}$ plus again the classical nonconformity estimator.

As already indicated, the proof of Theorem 4.7 is decomposed into several parts:

Lemma 4.8 (Local efficiency of the nonconformity estimator). *Let the assumptions of Theorem 4.7 be verified. Then* (4.30) *holds true.*

Proof. We have

$$\begin{aligned}
\eta_{\mathrm{NC},T}^2 &= \|u_h - \mathcal{I}_{\mathrm{Os}}(u_h)\|_{B,T}^2 \le \lambda_{M,T}\|\nabla(u_h - \mathcal{I}_{\mathrm{Os}}(u_h))\|_{0,T}^2 \\
&\le C\lambda_{M,T}\sum_{F\in\widetilde{\mathcal{F}}_T} h_F^{-1}\|[\![u_h]\!]\|_{0,F}^2 = C\sum_{F\in\widetilde{\mathcal{F}}_T}\frac{\lambda_{M,T}}{\alpha_F\gamma_{K,F}}\alpha_F\frac{\gamma_{K,F}}{h_F}\|[\![u_h]\!]\|_{0,F}^2 \qquad (4.32)\\
&\le \frac{C}{C_1}\left(\min_{F\in\widetilde{\mathcal{F}}_T}\alpha_F\right)^{-1}\frac{\lambda_{M,T}}{\lambda_{m,\widetilde{T}_T}}\sum_{F\in\widetilde{\mathcal{F}}_T}\alpha_F\frac{\gamma_{K,F}}{h_F}\|[\![u-u_h]\!]\|_{0,F}^2,
\end{aligned}$$

using Lemma 4.3, the lower bound in (4.22)–(4.23), and the fact that $[\![u_h - u]\!] = [\![u_h]\!]$. Recall from Lemma 4.3 that C depends only on d, k, and $\kappa_{\mathcal{T}}$. □

Lemma 4.9 (Norm estimate for the $\mathbf{RTN}_T^l$ space). *Let $\mathcal{T}_h$ be shape-regular. Then there exists a constant C, depending only on d, k, and $\kappa_{\mathcal{T}}$ such that for all $v_h \in \mathbf{RTN}_T^l$, there holds*

$$\|v_h\|_{0,T}^2 \le C\left\{h_T\sum_{F\in\mathcal{F}_T}\|v_h\cdot n_F\|_{0,F}^2 + \left(\sup_{r_h\in\mathbb{P}_{l-1}^d(T)}\frac{(v_h,r_h)_{0,T}}{\|r_h\|_{0,T}}\right)^2\right\}.$$

Proof. Use norm equivalence on finite-dimensional spaces, the Piola transformation, and scaling arguments. □

Lemma 4.10 (Local efficiency of the diffusive flux estimator). *Let the assumptions of Theorem 4.7 be verified. Then* (4.31) *holds true.*

Proof. Throughout this proof, let C denote a general constant not necessarily the same at each occurrence, depending only on d, k, and $\kappa_{\mathcal{T}}$. We put $v_h := (K\nabla u_h + \mathbf{t}_h)|_T \in \mathbf{RTN}_T^l$ and notice that, for $r_h \in \mathbb{P}_{l-1}^d(T)$,

$$\begin{aligned}(v_h, r_h)_{0,T} &= \theta \sum_{F\in\mathcal{F}_T} \omega_{T,F}(n_F^t K r_h, [\![u_h]\!])_{0,F} \\ &\le |\theta| C C_3 h_T^{-\frac{1}{2}} \|r_h\|_{0,T} \sum_{F\in\mathcal{F}_T} \gamma_{K,F} \|[\![u_h]\!]\|_{0,F},\end{aligned}$$

owing to the definition (4.15), the Cauchy–Schwarz inequality, the inverse inequality $\|r_h\|_{0,F} \le C h_T^{-1/2}\|r_h\|_{0,T}$ and the lower bound (4.24). Hence, using the definition (4.18) of $\eta_{\mathrm{DF},T}$, the previous lemma, the definition of $\mathbf{t}_h$ by (4.14)–(4.15), and the above inequality leads to

$$\begin{aligned}\eta_{\mathrm{DF},T}^2 &\le \frac{1}{\lambda_{m,T}}\|v_h\|_{0,T}^2 \le \frac{C}{\lambda_{m,T}}\left\{h_T \sum_{F\in\mathcal{F}_T}\|v_h\cdot n_F\|_{0,F}^2 + \left(\sup_{r_h\in\mathbb{P}_{l-1}^d(T)} \frac{(v_h, r_h)_{0,T}}{\|r_h\|_{0,T}}\right)^2\right\} \\ &\le \frac{C}{\lambda_{m,T}}\left\{h_T \sum_{F\in\mathcal{F}_T}\left\|\bar\omega_{T,F} n_T^t [\![K\nabla_h u_h]\!] + \alpha_F \frac{\gamma_{K,F}}{h_F}\Pi_l([\![u_h]\!])\right\|_{0,F}^2 \right. \\ &\qquad \left. + \theta^2 h_T^{-1} C_3^2 \sum_{F\in\mathcal{F}_T}\gamma_{K,F}^2 \|[\![u_h]\!]\|_{0,F}^2\right\}.\end{aligned}$$

We next study the three resulting terms separately. First of all,

$$\begin{aligned}&\frac{h_T}{\lambda_{m,T}}\sum_{F\in\mathcal{F}_T}\left\|\alpha_F\frac{\gamma_{K,F}}{h_F}\Pi_l([\![u_h]\!])\right\|_{0,F}^2 \le C\lambda_{m,T}^{-\frac{1}{2}}\sum_{F\in\mathcal{F}_T}\alpha_F\gamma_{K,F}\alpha_F\frac{\gamma_{K,F}}{h_F}\|\Pi_l([\![u_h]\!])\|_{0,F}^2 \\ \le\ & CC_2 \max_{F\in\mathcal{F}_T}\lambda_{m,T}^{-\frac{1}{2}}\alpha_F \sum_{F\in\mathcal{F}_T}\min\{\lambda_{M,T^-(F)}, \lambda_{M,T^+(F)}\}\alpha_F\frac{\gamma_{K,F}}{h_F}\|[\![u_h]\!]\|_{0,F}^2 \\ \le\ & CC_2 \max_{F\in\mathcal{F}_T}\alpha_F\frac{\lambda_{M,T}}{\lambda_{m,T}}\sum_{F\in\mathcal{F}_T}\alpha_F\frac{\gamma_{K,F}}{h_F}\|[\![u-u_h]\!]\|_{0,F}^2,\end{aligned}$$

where we have used (4.22) (the modification at the boundary has been skipped for simplicity). Similarly, for the second term we have

$$\theta^2\frac{1}{h_T\lambda_{m,T}}\sum_{F\in\mathcal{F}_T}\gamma_{K,F}^2\|[\![u_h]\!]\|_{0,F}^2 \le CC_3\theta^2\frac{\lambda_{M,T}}{\lambda_{m,T}}\left(\min_{F\in\tilde{\mathcal{F}}_T}\alpha_F\right)^{-1}\sum_{F\in\mathcal{F}_T}\alpha_F\frac{\gamma_{K,F}}{h_F}\|[\![u-u_h]\!]\|_{0,F}^2.$$

Finally,

$$\frac{h_T}{\lambda_{m,T}}\sum_{F\in\mathcal{F}_T}\bar\omega_{T,F}^2\left\|n_F^t[\![K\nabla_h u_h]\!]\right\|_{0,F}^2 \le CC_2C_3\max_{T'\in\mathcal{T}_T}\left\{\frac{\lambda_{M,T'}}{\lambda_{m,T'}}\right\}^2\|u-u_h\|_{B,\mathcal{T}_T}^2,$$

which can be deduced from [49, Proposition 3.2][5]. □

4.5 A posteriori error estimates for the reconstructed flux

The a posteriori error estimates of Section 4.3 are given for the DG approximate solution u_h, or, equivalently, taking into account the definition of the energy norm (4.5), for its flux $-K\nabla_h u_h$. In order to obtain them, we have used its $H(\mathrm{div},\Omega)$-conforming diffusive flux reconstruction $\mathbf{t}_h$. There arises a natural question whether we are also able to give an a posteriori error estimate for this (supposedly) improved flux, instead of the original estimate. Moreover, Theorem 4.6 indicates that the a posteriori error estimates of Section 4.3 will only lead to quasi-exactness with effectivity index $\sqrt{2}$. This is quite obvious because both u_h and $-K\nabla_h u_h$ are nonconforming (in the sense that $u_h \notin H_0^1(\Omega)$ and $-K\nabla_h u_h \notin H(\mathrm{div},\Omega)$) and consequently two estimators appear. We present here a possible remedy to this situation: since $\mathbf{t}_h \in H(\mathrm{div},\Omega)$ is such that $\nabla\cdot\mathbf{t}_h = \Pi_l(f)$, it suffices to use the results of [103] in order to obtain the same estimates as for mixed finite elements:

Theorem 4.11 (Abstract a posteriori error estimate for the reconstructed flux)**.** *Let u be the unique solution of (4.6), let u_h be its discontinuous Galerkin approximation given by (4.10), and let $\mathbf{t}_h$ be its diffusive flux reconstruction given by (4.14)–(4.15). Then*

$$\|K^{-\frac{1}{2}}\mathbf{t}_h + K^{\frac{1}{2}}\nabla u\|_{0,\Omega}^2 \leq \inf_{s\in H_0^1(\Omega)} \|K^{-\frac{1}{2}}\mathbf{t}_h + K^{\frac{1}{2}}\nabla s\|_{0,\Omega}^2 + \sum_{T\in\mathcal{T}_h} m_{T,K}^2 \|f-\Pi_l(f)\|_{0,T}^2.$$

For practical purposes, a first choice for s in the above theorem is $s = \mathcal{I}_{\mathrm{Os}}(u_h)$. However, more precise reconstructions are suggested and studied in [103] for mixed finite elements.

Concerning the efficiency of this framework, we have the following result, which in contrast to Section 4.4.1 gives full asymptotic exactness (*i.e.*, effectivity index equal to 1, up to the residual (or data oscillation) term).

Theorem 4.12 (Global efficiency of the abstract estimate for the reconstructed flux)**.** *Let u be the unique solution of (4.6), let u_h be its discontinuous Galerkin approximation given by (4.10), and let $\mathbf{t}_h$ be its diffusive flux reconstruction given by (4.14)–(4.15). Let the a posteriori error estimate for $\mathbf{t}_h$ be given by Theorem 4.11. Then*

$$\begin{aligned}&\inf_{s\in H_0^1(\Omega)} \|K^{-\frac{1}{2}}\mathbf{t}_h + K^{\frac{1}{2}}\nabla s\|_{0,\Omega}^2 + \sum_{T\in\mathcal{T}_h} m_{T,K}^2 \|f-\Pi_l(f)\|_{0,T}^2 \\ &\leq \|K^{-\frac{1}{2}}\mathbf{t}_h + K^{\frac{1}{2}}\nabla u\|_{0,\Omega}^2 + \sum_{T\in\mathcal{T}_h} m_{T,K}^2 \|f-\Pi_l(f)\|_{0,T}^2.\end{aligned}$$

[5] Chapitre 3, Proposition 3.5

4.6 Improved energy norm a posteriori error estimates in the general case

We present in this section an extension of our analysis of Section 4.3 to the general case (4.1a)–(4.1b). Again at this stage, neither assumptions on the data other than those stated in Section 4.2.2, nor the mesh shape-regularity, are needed.

4.6.1 Abstract framework

The following general abstract framework has been proved in [102, Lemma 7.1].

Lemma 4.13 (Abstract framework in the general case)**.** *Let $u \in H_0^1(\Omega)$ and $u_h \in H^1(\mathcal{T}_h)$ be arbitrary. Then*

$$\begin{aligned}\|u-u_h\|_B \;\le\; \inf_{s\in H_0^1(\Omega)}\Big\{\|u_h-s\|_B + \sup_{\varphi\in H_0^1(\Omega),\,\|\varphi\|_B=1}\big(B(u-u_h,\varphi)\\ + \big(\beta\cdot\nabla_h(u_h-s)+\tfrac12(\nabla\cdot\beta)(u_h-s),\varphi\big)\big)\Big\}.\end{aligned}$$

Remark (Comparison with the abstract framework of Lemma 4.1). In comparison with the abstract framework of Lemma 4.1, Lemma 4.13 is applicable to the general advection–diffusion–reaction case. In particular, there is an additional contribution from the non-symmetric part of the bilinear form $B(\cdot,\cdot)$, which can be evaluated using an arbitrary $s \in H_0^1(\Omega)$. The price for this generality is that Lemma 4.13 yields a triangular-like inequality instead of a Pythagorean-like inequality.

4.6.2 Abstract a posteriori error estimate

An abstract form of our a posteriori error estimate now takes the following form (compare with Theorem 4.2).

Theorem 4.14 (Abstract a posteriori error estimate in the general case)**.** *Let u be the unique solution of (4.6) and let $u_h \in H^1(\mathcal{T}_h)$ be arbitrary. Then*

$$\begin{aligned}\|u-u_h\|_B \;\le\; \inf_{s\in H_0^1(\Omega)}\Big\{&\|u_h-s\|_B\\ &+\inf_{\mathbf{t}\in H(\mathrm{div},\Omega)}\sup_{\varphi\in H_0^1(\Omega),\,\|\varphi\|_B=1}\big((f-\nabla\cdot\mathbf{t}-\beta\cdot\nabla s-\mu s,\varphi)\\ &-(K\nabla_h u_h+\mathbf{t},\nabla\varphi)+\big((\mu-\tfrac12\nabla\cdot\beta)(s-u_h),\varphi\big)\big)\Big\}.\end{aligned}\tag{4.33}$$

Proof. We use (4.6) in Lemma 4.13, keep $s \in H_0^1(\Omega)$ arbitrary, introduce an arbitrary $\mathbf{t} \in H(\mathrm{div},\Omega)$, and employ the Green theorem to infer

$$\begin{aligned}&B(u-u_h,\varphi) + (\beta\cdot\nabla_h(u_h-s) + \tfrac{1}{2}(\nabla\cdot\beta)(u_h-s),\varphi) = (f - \nabla\cdot\mathbf{t} - \beta\cdot\nabla_h u_h - \mu u_h,\varphi)\\ &-(K\nabla_h u_h + \mathbf{t},\nabla\varphi) + \big(\beta\cdot\nabla_h(u_h-s) + \tfrac{1}{2}(\nabla\cdot\beta)(u_h-s),\varphi\big),\end{aligned}$$

whence the assertion of the theorem follows easily. □

Remark (A computable a posteriori error estimate). Similarly as in the pure diffusion case (*cf.* Remark 4.3.2), it follows readily from (4.33), using the Friedrichs inequality $\|\varphi\|^2 \le C_{\mathrm{F},\Omega} h_\Omega^2 \|\nabla\varphi\|^2$ or the inequality $\|\varphi\| \le \min_{T\in\mathcal{T}_h}\{\tilde\mu_{m,T}^{1/2}\}^{-1}\big\|\big(\mu - \frac{1}{2}\nabla\cdot\beta\big)^{1/2}\varphi\big\|$, the Cauchy–Schwarz inequality, the definition (4.5) of the energy semi-norm, and the fact that $\|\varphi\|_B = 1$,

$$\begin{aligned}\|u-u_h\|_B \;\le\;& \|u_h - s\|_B\\ &+ \min\left\{\frac{C_{\mathrm{F},\Omega}^{1/2} h_\Omega}{\min_{T\in\mathcal{T}_h}\lambda_{m,T}^{1/2}}, \frac{1}{\min_{T\in\mathcal{T}_h}\tilde\mu_{m,T}^{1/2}}\right\}\|f - \nabla\cdot\mathbf{t} - \beta\cdot\nabla s - \mu s\|_{0,\Omega}\\ &+ \Big(\|K^{\frac{1}{2}}\nabla_h u_h + K^{-\frac{1}{2}}\mathbf{t}_h\|_{0,\Omega}^2 + \|\big(\mu - \tfrac{1}{2}\nabla\cdot\beta\big)^{\frac{1}{2}}(u_h - s)\|_{0,\Omega}^2\Big)^{1/2}\end{aligned}$$

for any $s \in H_0^1(\Omega)$ and any $\mathbf{t} \in H(\mathrm{div},\Omega)$. Again, this is a fully computable and scheme-independent estimate, but all the points from Remark 4.3.2 apply here as well.

Remark (Another form of Theorem 4.14). The estimate of Theorem 4.14 can be changed into

$$\begin{aligned}\|u-u_h\|_B \le \inf_{s\in H_0^1(\Omega)}\Bigg\{&\|u_h - s\|_B \\ &+ \inf_{\mathbf{q}\in H(\mathrm{div},\Omega)}\inf_{\mathbf{t}\in H(\mathrm{div},\Omega)}\sup_{\varphi\in H_0^1(\Omega),\,\|\varphi\|_B=1}\Big|(f - \nabla\cdot\mathbf{t} - \nabla\cdot\mathbf{q} - (\mu - \nabla\cdot\beta)u_h,\varphi)\\ &- (K\nabla_h u_h + \mathbf{t},\nabla\varphi) + (\nabla\cdot\mathbf{q} - \nabla\cdot(\beta s),\varphi) - \big(\tfrac{1}{2}(\nabla\cdot\beta)(u_h - s),\varphi\big)\Big|\Bigg\}.\end{aligned} \tag{4.34}$$

Again, Theorem 4.14 gives a framework for a quasi-optimal a posteriori error estimate (see Section 4.7.1 below), which is however not practically computable. In the next sections, we present its locally computable version. To this purpose, we would like to keep the choice of the $H_0^1(\Omega)$-conforming scalar function s and of the $H(\mathrm{div},\Omega)$-conforming diffusive flux $\mathbf{t}$ the same as in Sections 4.3.3 and 4.3.4, respectively. With this choice, however, the residual $f - \nabla\cdot\mathbf{t} - \beta\cdot\nabla s - \mu s$ does not satisfy an orthogonality property as (4.16). It is not

even of zero mean necessarily, which would be necessary to obtain a computable estimate as in (4.21). In order to recover (at least partially) these properties, we will employ the form of Theorem 4.14 given by (4.34), where $\mathbf{q}$ will be a suitable $H(\mathrm{div},\Omega)$-conforming convective flux reconstruction.

4.6.3 Convective flux reconstruction

Our $H(\mathrm{div},\Omega)$-conforming convective flux reconstruction $\mathbf{q}_h$ will belong to $\mathbf{RTN}_h^l$ with $l=k$ or $l=k-1$ and we prescribe it locally on all $T\in\mathcal{T}_h$, as follows:

$$(\mathbf{q}_h{\cdot}n_F, q_h)_{0,F} = (\beta{\cdot}n_F\{u_h\}+\gamma_{\beta,F}[\![u_h]\!], q_h)_{0,F} \qquad \forall q_h\in\mathbb{P}_l(F),\ \forall F\in\mathcal{F}_T, \tag{4.35}$$

$$(\mathbf{q}_h, r_h)_{0,T} = (u_h, \beta{\cdot}r_h)_{0,T} \qquad \forall r_h\in\mathbb{P}_{l-1}^d(T). \tag{4.36}$$

Note in particular that the quantities prescribing the moments of $\mathbf{q}_h{\cdot}n_F$ are uniquely defined for each face $F\in\mathcal{F}_h$, whence the continuity of the normal trace of $\mathbf{q}_h$. By this construction, we have the following generalization of Lemma 4.4:

Lemma 4.15 (Reconstructed advection–diffusion–reaction residual)**.** *There holds*

$$(\nabla{\cdot}\mathbf{t}_h + \nabla{\cdot}\mathbf{q}_h + (\mu-\nabla{\cdot}\beta)u_h, \xi_h)_{0,T} = (f,\xi_h)_{0,T} \qquad \forall T\in\mathcal{T}_h,\ \forall\xi_h\in\mathbb{P}_l(T).$$

Moreover, using that $\nabla{\cdot}\mathbf{t}_h|_T\in\mathbb{P}_l(T)$ *and* $\nabla{\cdot}\mathbf{q}_h|_T\in\mathbb{P}_l(T)$ *for all* $T\in\mathcal{T}_h$*,*

$$(\nabla{\cdot}\mathbf{t}_h + \nabla{\cdot}\mathbf{q}_h + \Pi_l((\mu-\nabla{\cdot}\beta)u_h))|_T = \Pi_l(f)|_T \qquad \forall T\in\mathcal{T}_h,$$

and, when in particular μ *and* $\nabla{\cdot}\beta$ *are elementwise constant and when* $l=k$ *in the diffusive and convective flux reconstructions,*

$$(\nabla{\cdot}\mathbf{t}_h + \nabla{\cdot}\mathbf{q}_h + (\mu-\nabla{\cdot}\beta)u_h)|_T = \Pi_k(f)|_T \qquad \forall T\in\mathcal{T}_h.$$

Proof. Let $\xi_h\in\mathbb{P}_l(T)$ be arbitrary. Owing to the Green theorem, the fact that $\xi_h|_F\in\mathbb{P}_l(F)$ for all $F\in\mathcal{F}_T$, $\nabla\xi_h\in\mathbb{P}_{l-1}^d(T)$, the definitions (4.14)–(4.15) of $\mathbf{t}_h$ and the definitions (4.35)–

(4.36) of $\mathbf{q}_h$, respectively, and putting $v_h = \xi_h$ on T and $v_h = 0$ otherwise, we infer

$$\begin{aligned}
& (\nabla\cdot\mathbf{t}_h + \nabla\cdot\mathbf{q}_h + (\mu - \nabla\cdot\beta)u_h, \xi_h)_{0,T} \\
= \; & -(\mathbf{t}_h, \nabla\xi_h)_{0,T} + \sum_{F\in\mathcal{F}_T} (\mathbf{t}_h\cdot n_T, \xi_h)_{0,F} - (\mathbf{q}_h, \nabla\xi_h)_{0,T} + \sum_{F\in\mathcal{F}_T} (\mathbf{q}_h\cdot n_T, \xi_h)_{0,F} \\
& + ((\mu - \nabla\cdot\beta)u_h, \xi_h)_{0,T} = (K\nabla_h u_h, \nabla_h v_h)_{0,T} - (u_h, \beta\cdot\nabla_h v_h)_{0,T} \\
& - \sum_{F\in\mathcal{F}_T} \Big\{ \theta(n_F^t\{K\nabla_h v_h\}_\omega, [\![u_h]\!])_{0,F} + \Big(n_F^t\{K\nabla_h u_h\}_\omega - \alpha_F \frac{\gamma_{K,F}}{h_F}[\![u_h]\!], [\![v_h]\!]\Big)_{0,F} \Big\} \\
& + \sum_{F\in\mathcal{F}_T} (\beta\cdot n_F\{u_h\} + \gamma_{\beta,F}[\![u_h]\!], [\![v_h]\!])_{0,F} + ((\mu - \nabla\cdot\beta)u_h, v_h)_{0,T} \\
= \; & B_h(u_h, v_h) = (f, v_h) = (f, \xi_h)_{0,T},
\end{aligned}$$

employing finally the definition of the DG bilinear form (4.9) and that of the DG approximate solution (4.10). □

4.6.4 Locally computable a posteriori error estimate

We are now ready to state our practical locally computable a posteriori error estimate for DG methods and the problem (4.1a)–(4.1b).

While we keep the definitions of the nonconformity estimator $\eta_{\mathrm{NC},T}$ (4.17) and that of the diffusive flux estimator $\eta_{\mathrm{DF},T}$ (4.18) as in Section 4.3.5, the *residual estimator* $\eta_{\mathrm{R},T}$ will now be defined by

$$\eta_{\mathrm{R},T} := m_{T,K,\beta,\mu}\|f - \nabla\cdot\mathbf{t}_h - \nabla\cdot\mathbf{q}_h - (\mu - \nabla\cdot\beta)u_h\|_{0,T}, \tag{4.37}$$

where

$$m_{T,K,\beta,\mu}^2 := \min\left\{ C_{\mathrm{P}}\frac{h_T^2}{\lambda_{m,T}}, \frac{1}{\tilde{\mu}_{m,T}} \right\}$$

for all $T \in \mathcal{T}_h$; recall that C_{P} is the constant from the Poincaré inequality (4.19). We next define two *advection estimators* $\eta_{\mathrm{C},1,T}$ and $\eta_{\mathrm{C},2,T}$ respectively by

$$\eta_{\mathrm{C},1,T} := m_{T,K,\beta,\mu}\|\nabla\cdot(\mathbf{q}_h - \beta s_h) - \Pi_0(\nabla\cdot(\mathbf{q}_h - \beta s_h))\|_{0,T} \tag{4.38}$$

and

$$\eta_{\mathrm{C},2,T} := \frac{1}{\tilde{\mu}_{m,T}^{1/2}}\big\|\tfrac{1}{2}(\nabla\cdot\beta)(u_h - s_h)\big\|_{0,T}, \tag{4.39}$$

with $s_h = \mathcal{I}_{\mathrm{Os}}(u_h)$. Finally, let

$$m_{F,K,\beta,\mu}^2 := \min\left\{ \max_{T;\,F\in\mathcal{F}_T}\left\{ C_{\mathrm{F},T,F}\frac{|F|h_T^2}{|T|\lambda_{m,T}} \right\}, \max_{T;\,F\in\mathcal{F}_T}\left\{ \frac{|F|}{|T|\tilde{\mu}_{m,T}} \right\} \right\} \tag{4.40}$$

for a face $F \in \mathcal{F}_h$. Here $C_{\mathrm{F},T,F}$ is the constant from the generalized Friedrichs inequality, which states that

$$\|\varphi - \Pi_{0,F}(\varphi)\|_{0,T}^2 \le C_{\mathrm{F},T,F} h_T^2 \|\nabla\varphi\|_{0,T}^2, \tag{4.41}$$

where $\Pi_{l,F}$ is the L^2-orthogonal projection onto piecewise polynomials of degree l on the face F. It follows from [100, Lemma 4.1] that $C_{\mathrm{F},T,F} = 3d$ for a simplex T and its face F. The *upwinding estimator* $\eta_{\mathrm{U},T}$ is defined by

$$\eta_{\mathrm{U},T} := \sum_{F\in\mathcal{F}_T} m_{F,K,\beta,\mu} \|\Pi_{0,F}((\mathbf{q}_h - \beta s_h)\cdot n_F)\|_{0,F}. \tag{4.42}$$

We then have the following a posteriori error estimate:

Theorem 4.16 (Locally computable a posteriori error estimate in the general case)**.** *Let u be the unique solution of* (4.6) *and let u_h be its discontinuous Galerkin approximation given by* (4.10)*. Then*

$$\begin{aligned}\|u - u_h\|_B \;\le\; & \left\{\sum_{T\in\mathcal{T}_h} \eta_{\mathrm{NC},T}^2\right\}^{1/2} \\ & + \left\{\sum_{T\in\mathcal{T}_h} \Big(\eta_{\mathrm{R},T} + (\eta_{\mathrm{DF},T}^2 + \eta_{\mathrm{C},2,T}^2)^{1/2} + \eta_{\mathrm{C},1,T} + \eta_{\mathrm{U},T}\Big)^2\right\}^{1/2}.\end{aligned}$$

Proof. We start by putting $s = s_h = \mathcal{I}_{\mathrm{Os}}(u_h)$, $\mathbf{t} = \mathbf{t}_h$, and $\mathbf{q} = \mathbf{q}_h$ in the abstract estimate of (4.34). We next write

$$\begin{aligned}&(f - \nabla\cdot\mathbf{t}_h - \nabla\cdot\mathbf{q}_h - (\mu - \nabla\cdot\beta)u_h, \varphi) - (K\nabla_h u_h + \mathbf{t}_h, \nabla\varphi) + (\nabla\cdot\mathbf{q}_h - \nabla\cdot(\beta s_h), \varphi) \\ &-\big(\tfrac{1}{2}(\nabla\cdot\beta)(u_h - s_h), \varphi\big) = \sum_{T\in\mathcal{T}_h}\Big\{(f - \nabla\cdot\mathbf{t}_h - \nabla\cdot\mathbf{q}_h - (\mu - \nabla\cdot\beta)u_h, \varphi - \Pi_0(\varphi))_{0,T} \\ &-(K\nabla u_h + \mathbf{t}_h, \nabla\varphi)_{0,T} - \big(\tfrac{1}{2}(\nabla\cdot\beta)(u_h - s_h), \varphi\big)_{0,T} + (\nabla\cdot(\mathbf{q}_h - \beta s_h), \varphi - \Pi_0(\varphi))_{0,T} \\ &+ \sum_{F\in\mathcal{F}_T}((\mathbf{q}_h - \beta s_h)\cdot n_T, \Pi_0(\varphi))_{0,F}\Big\},\end{aligned}$$

using Lemma 4.15 in the first term and subtracting $(\nabla\cdot(\mathbf{q}_h - \beta s_h), \Pi_0(\varphi))_{0,T}$ and adding the same quantity rewritten using the Green theorem in the last but one term. Next note that in this last term, we may replace $\nabla\cdot(\mathbf{q}_h - \beta s_h)$ by $\nabla\cdot(\mathbf{q}_h - \beta s_h) - \Pi_0(\nabla\cdot(\mathbf{q}_h - \beta s_h))$, and similarly in the last term, we may replace $(\mathbf{q}_h - \beta s_h)\cdot n_T$ by $\Pi_{0,F}((\mathbf{q}_h - \beta s_h)\cdot n_T)$. The above expression is thus bounded, using the Cauchy–Schwarz inequality, the inequality

$$\|\varphi - \Pi_0(\varphi)\|_{0,T} \le m_{T,K,\beta,\mu}\|\varphi\|_{B,T},$$

which follows from the Poincaré inequality (4.19) and from the definition of the energy norm (4.5), and finally using [101, Lemma 4.5] for the last term, by

$$\sum_{T\in\mathcal{T}_h}\Big(\eta_{\mathrm{R},T}+(\eta_{\mathrm{DF},T}^2+\eta_{\mathrm{C},2,T}^2)^{1/2}+\eta_{\mathrm{C},1,T}+\eta_{\mathrm{U},T}\Big)\|\varphi\|_{B,T}.$$

Using the Cauchy–Schwarz inequality, noticing that $\|\varphi\|_B=1$, and adding the nonconformity estimator, which appears directly as the first term in (4.34), concludes the proof. □

Remark (Properties of the estimate of Theorem 4.16). As in the pure diffusion case, we remark that the estimate of Theorem 4.16 yields a guaranteed upper bound, the residual represents a higher-order term, neither C_{P} nor $C_{\mathrm{F},T,F}$ depend on the polynomial degree of u_h, whence the estimate is valid uniformly with respect to k, no polynomial data form is needed at this stage, and, finally, the estimate is valid for any $\mathbf{t}_h,\mathbf{q}_h\in H(\mathrm{div},\Omega)$ such that $(\nabla{\cdot}\mathbf{t}_h+\nabla{\cdot}\mathbf{q}_h+(\mu-\nabla{\cdot}\beta)u_h,1)_{0,T}=(f,1)_{0,T}$ for all $T\in\mathcal{T}_h$.

Remark (Mean values in $\eta_{\mathrm{C},1,T}$ and $\eta_{\mathrm{U},T}$). Since $\|g-\Pi_0(g)\|_{0,\Omega}\le\|g\|_{0,\Omega}$ and $\|\Pi_0(g)\|_{0,\Omega}\le\|g\|_{0,\Omega}$, where $\|\cdot\|_{0,\Omega}$ denotes the L^2-norm for a square-integrable function g, the estimators $\eta_{\mathrm{C},1,T}$ and $\eta_{\mathrm{U},T}$ may be considerably smaller in the advection-dominated case when compared to the situation where the piecewise constant projection is not subtracted/used.

4.7 Efficiency of the estimates in the general case

In this section, we first rapidly check the (global) efficiency of the abstract framework of Theorem 4.14. We then investigate in detail the (local) efficiency of the a posteriori error estimate of Theorem 4.16.

4.7.1 Global efficiency of the abstract estimate

Theorem 4.17 (Global efficiency of the abstract estimate in the general case)**.** *Let u be the unique solution of (4.6) and let $u_h\in H^1(\mathcal{T}_h)$ be arbitrary. Let the a posteriori error estimate be given by Theorem 4.14. Then*

$$\begin{aligned}&\inf_{s\in H_0^1(\Omega)}\Big\{\|u_h-s\|_B+\inf_{\mathbf{t}\in H(\mathrm{div},\Omega)}\sup_{\varphi\in H_0^1(\Omega),\,\|\varphi\|_B=1}\big((f-\nabla{\cdot}\mathbf{t}-\beta{\cdot}\nabla s-\mu s,\varphi)\\&\quad-(K\nabla_h u_h+\mathbf{t},\nabla\varphi)+\big((\mu-\tfrac12\nabla{\cdot}\beta)(s-u_h),\varphi\big)\big)\Big\}\\\le\;&2\|u-u_h\|_B.\end{aligned}$$

Proof. It suffices to put $s = u$ and $\mathbf{t} = -K\nabla u$ in Theorem 4.14 and to use the Cauchy–Schwarz inequality and the fact that $\|\varphi\|_B = 1$. □

Before we start to investigate the local efficiency of the locally computable estimate, we note that the same global efficiency result holds true also for the estimate (4.34) – it suffices to put in addition $\mathbf{q} = \beta u$. In addition, similar observations to those given after Theorem 4.6 hold true also in the present case.

4.7.2 Local efficiency of the locally computable estimate

The following theorem is a generalization of Theorem 4.7:

Theorem 4.18 (Local efficiency of the locally computable estimate in the general case)**.** *Let $\mathcal{T}_h$ be shape-regular, let f be a piecewise polynomial of degree m, and let, for the sake of simplicity, $\nabla\cdot(\mathbf{q}_h - \beta s_h) \in \mathbb{P}_l$ on all $T \in \mathcal{T}_h$. Let next u be theu unique solution of* (4.6) *and let u_h be its discontinuous Galerkin approximation given by* (4.10)*. Assume* (4.22)–(4.24) *for the weights $\omega_{T,F}$ and for the penalty parameters $\gamma_{K,F}$ and that $\gamma_{\beta,F} \le \|\beta\|_{\infty,T}$ for all $T \in \mathcal{T}_h$ and $F \in \mathcal{F}_T$ (both $\gamma_{K,F}$ and $\gamma_{\beta,F}$ are for the simplicity supposed facewise constant). Let finally the a posteriori error estimate be given by Theorem 4.16, with in particular $\eta_{\mathrm{NC},T}$ given by* (4.17)*, $\eta_{\mathrm{R},T}$ by* (4.37)*, $\eta_{\mathrm{DF},T}$ by* (4.18)*, $\eta_{\mathrm{C},1,T}$ by* (4.38)*, $\eta_{\mathrm{C},2,T}$ by* (4.39)*, and $\eta_{\mathrm{U},T}$ by* (4.42)*. Recall the notation $\lambda_{m,\widetilde{\mathcal{T}}_T}$ from* (4.28) *and put*

$$\tilde{\mu}_{m,\mathcal{T}_T} := \min_{T' \in \mathcal{T}_T} \tilde{\mu}_{m,T'}, \quad c_{\beta,\widetilde{\mathcal{F}}_T} := \min_{F \in \widetilde{\mathcal{F}}_T} \gamma_{\beta,F}, \tag{4.43}$$

and $R(u_h) := f + \nabla\cdot(K\nabla_h u_h) - \beta\cdot\nabla_h u_h - \mu u_h$. *Then, for each* $T \in \mathcal{T}_h$, *there holds*

$$\eta_{\mathrm{NC},T} \leq C\left(\lambda_{M,T}^{1/2}h_T^{-1} + \left\|\mu - \tfrac{1}{2}\nabla\cdot\beta\right\|_{\infty,T}^{1/2}\right)\min\left\{\frac{h_T}{\lambda_{m,\widetilde{\mathcal{T}}_T}^{1/2}}, \frac{h_T^{1/2}}{c_{\beta,\widetilde{\mathcal{F}}_T}^{1/2}}\right\}\|u-u_h\|_{B,*,\widetilde{\mathcal{F}}_T}, \tag{4.44}$$

$$\eta_{\mathrm{C},2,T} \leq C\frac{\left|\frac{1}{2}\nabla\cdot\beta\right|}{\tilde{\mu}_{m,T}^{1/2}}\min\left\{\frac{h_T}{\lambda_{m,\widetilde{\mathcal{T}}_T}^{1/2}}, \frac{h_T^{1/2}}{c_{\beta,\widetilde{\mathcal{F}}_T}^{1/2}}\right\}\|u-u_h\|_{B,*,\widetilde{\mathcal{F}}_T}, \tag{4.45}$$

$$\eta_{\mathrm{U},T} \leq C\min\left\{\frac{1}{\lambda_{m,\mathcal{T}_T}^{1/2}}, \frac{1}{h_T\tilde{\mu}_{m,\mathcal{T}_T}^{1/2}}\right\}\|\beta\|_{\infty,T}\min\left\{\frac{h_T}{\lambda_{m,\widetilde{\mathcal{T}}_T}^{1/2}}, \frac{h_T^{1/2}}{c_{\beta,\widetilde{\mathcal{F}}_T}^{1/2}}\right\}\|u-u_h\|_{B,*,\widetilde{\mathcal{F}}_T}, \tag{4.46}$$

$$\eta_{\mathrm{C},1,T} \leq C\min\left\{\frac{1}{\lambda_{m,\mathcal{T}_T}^{1/2}}, \frac{1}{h_T\tilde{\mu}_{m,\mathcal{T}_T}^{1/2}}\right\}\|\beta\|_{\infty,T}\min\left\{\frac{h_T}{\lambda_{m,\widetilde{\mathcal{T}}_T}^{1/2}}, \frac{h_T^{1/2}}{c_{\beta,\widetilde{\mathcal{F}}_T}^{1/2}}\right\}\|u-u_h\|_{B,*,\widetilde{\mathcal{F}}_T}, \tag{4.47}$$

$$\begin{aligned}\eta_{\mathrm{DF},T} \leq{}& \tilde{C}\left(\frac{\lambda_{M,T}}{\lambda_{m,T}}\right)^{\frac{1}{2}}\|u-u_h\|_{B,*,\mathcal{F}_T} + \tilde{C}\max_{T'\in\mathcal{T}_T}\left\{\frac{\lambda_{M,T}}{\lambda_{m,T}}\right\}^{1/2}\\ &\left(\sum_{T'\in\mathcal{T}_T}\frac{h_{T'}}{\lambda_{m,T'}^{1/2}}\|R(u_h)\|_{0,T'} + \sum_{T'\in\mathcal{T}_T}\left\{\max\left\{1, \frac{\|\mu\|_{\infty,T'}}{\tilde{\mu}_{m,T'}}\right\}\right.\right.\\ &\left.\left.+\left(1+\frac{\|\mu-\frac{1}{2}\nabla\cdot\beta\|_{\infty,T'}^{1/2}}{\lambda_{m,T'}^{1/2}}h_{T'}\right) + \frac{\|\beta\|_{\infty,T'}}{\lambda_{m,T'}}h_{T'}\right\}\|u-u_h\|_{B,\mathcal{T}_T}\right),\end{aligned} \tag{4.48}$$

and

$$\begin{aligned}\|R(u_h)\|_{0,T} \leq{}& \bar{C}\left(\frac{\lambda_{M,T}^{1/2}}{h_T} + \min\left\{\frac{\|\mu\|_{\infty,T}}{\tilde{\mu}_{m,T}^{1/2}} + \frac{\|\beta\|_{\infty,T}}{\lambda_{m,T}^{1/2}}, \frac{\|\mu-\nabla\cdot\beta\|_{\infty,T}}{\tilde{\mu}_{m,T}^{1/2}} + \frac{\|\beta\|_{\infty,T}}{\tilde{\mu}_{m,T}^{1/2}h_T}\right\}\right)\\ &\times\|u-u_h\|_{B,T}.\end{aligned} \tag{4.49}$$

Here, the constants C *depend only on the space dimension* d, *on the maximal polynomial degree* k, *on the shape regularity parameter* $\kappa_{\mathcal{T}}$, *on the DG parameters* α_F, *and on the constant* C_1 *from* (4.22)–(4.23), $\tilde{C}$ *in addition depends on the polynomial degree* m *of* f, *on the DG parameter* θ, *and on the constant* C_3 *from* (4.24), *and* $\bar{C}$ *depends only on* d, k, m, *and* $\kappa_{\mathcal{T}}$.

Proof. Since $\eta_{\mathrm{R},T}$ is a higher-order term owing to Lemma 4.15 (given by $m_{T,K,\beta,\mu}\,\|f - \Pi_k(f)\|_{0,T}$ when μ and $\nabla\cdot\beta$ are elementwise constant and when $l = k$), we prove only the efficiency of the other estimates. In this proof, we denote by C a general constant not

necessarily the same at each occurrence, depending only on d, k, and $\kappa_{\mathcal{T}}$. Also, recall that it follows from (4.32) that

$$\sum_{F\in\widetilde{\mathcal{F}}_T} h_F^{-1}\|[\![u_h]\!]\|_{0,F}^2 \le C_1^{-1}(\min_{F\in\widetilde{\mathcal{F}}_T}\alpha_F)^{-1}\frac{1}{\lambda_{m,\widetilde{\mathcal{I}}_T}}\|u-u_h\|_{B,*,\widetilde{\mathcal{F}}_T}^2. \tag{4.50}$$

Similarly, one obviously has, using (4.29) and (4.43),

$$\sum_{F\in\widetilde{\mathcal{F}}_T} \|[\![u_h]\!]\|_{0,F}^2 \le \frac{1}{c_{\beta,\widetilde{\mathcal{F}}_T}}\|u-u_h\|_{B,*,\widetilde{\mathcal{F}}_T}^2. \tag{4.51}$$

We begin with $\eta_{\mathrm{NC},T}$. As the estimate for its diffusive part is given by (4.32), we only estimate

$$\left\|\left(\mu-\tfrac{1}{2}\nabla\cdot\beta\right)^{\frac{1}{2}}(u_h-\mathcal{I}_{\mathrm{Os}}(u_h))\right\|_{0,T}^2 \le C\left\|\mu-\tfrac{1}{2}\nabla\cdot\beta\right\|_{\infty,T}\sum_{F\in\widetilde{\mathcal{F}}_T} h_F\|[\![u_h]\!]\|_{0,F}^2,$$

using Lemma 4.3. We can further bound this term either by (4.50) or by (4.51), whence (4.44) follows.

Similarly, (4.45) is readily deduced, using

$$\eta_{\mathrm{C},2,T}^2 \le C\frac{\left|\frac{1}{2}\nabla\cdot\beta\right|^2}{\tilde{\mu}_{m,T}}\sum_{F\in\widetilde{\mathcal{F}}_T} h_F\|[\![u_h]\!]\|_{0,F}^2.$$

Next, we remark that, for all $F\in\mathcal{F}_T$,

$$m_{F,K,\beta,\mu}^2 \le C\min\left\{\frac{h_T}{\lambda_{M,\mathcal{T}_T}},\frac{1}{h_T\,\tilde{\mu}_{m,\mathcal{T}_T}}\right\},$$

whence, employing (4.35) and the fact that $\|\Pi_{0,F}(g)\|_{0,F}\le\|g\|_{0,F}$, we get, with $s_h=\mathcal{I}_{\mathrm{Os}}(u_h)$,

$$\begin{aligned}
&\eta_{\mathrm{U},T}\\
\le\;& C\min\left\{\frac{h_T^{1/2}}{{\lambda_{M,\mathcal{T}_T}}^{1/2}},\frac{1}{h_T^{1/2}\,\tilde{\mu}_{m,\mathcal{T}_T}^{1/2}}\right\}\sum_{F\in\mathcal{F}_T}\|\beta\cdot n_F\{u_h\}+\gamma_{\beta,F}[\![u_h]\!]-\beta\cdot n_F s_h\|_{0,F}\\
\le\;& C\min\left\{\frac{h_T^{1/2}}{{\lambda_{M,\mathcal{T}_T}}^{1/2}},\frac{1}{h_T^{1/2}\,\tilde{\mu}_{m,\mathcal{T}_T}^{1/2}}\right\}\sum_{F\in\mathcal{F}_T}\left\{\|\gamma_{\beta,F}[\![u_h]\!]\|_{0,F}\right.\\
&\left.+\tfrac{1}{2}\sum_{T';F\in\mathcal{F}_{T'}}\|\beta\cdot n_F(u_h-s_h)\|_{0,F}\right\}\\
\le\;& C\min\left\{\frac{1}{{\lambda_{M,\mathcal{T}_T}}^{1/2}},\frac{1}{h_T\tilde{\mu}_{m,\mathcal{T}_T}^{1/2}}\right\}\|\beta\|_{\infty,T}\sum_{F\in\widetilde{\mathcal{F}}_T} h_F^{1/2}\|[\![u_h]\!]\|_{0,F},
\end{aligned}$$

using also the inequality

$$\|u_h - \mathcal{I}_{\rm Os}(u_h)\|_{0,F} \le C \sum_{F';F'\cap F\neq\emptyset} \|[\![u_h]\!]\|_{0,F'}$$

valid for the Oswald interpolation operator.

We next prove the efficiency of $\eta_{{\rm C},1,T}$. First of all,

$$\begin{aligned} & m_{T,K,\beta,\mu}\|\nabla\cdot(\mathbf{q}_h - \beta s_h) - \Pi_0(\nabla\cdot(\mathbf{q}_h - \beta s_h))\|_{0,T} \\ \le\ & m_{T,K,\beta,\mu}\|\nabla\cdot(\mathbf{q}_h - \beta s_h)\|_{0,T} = m_{T,K,\beta,\mu} \sup_{\xi_h\in\mathbb{P}_l(T)} \frac{(\nabla\cdot(\mathbf{q}_h - \beta s_h),\xi_h)_{0,T}}{\|\xi_h\|_{0,T}}, \end{aligned}$$

using the assumption $\nabla\cdot(\mathbf{q}_h - \beta s_h) \in \mathbb{P}_l$ on all $T \in \mathcal{T}_h$. Next, using the Green theorem, the definition of $\mathbf{q}_h$ (4.35)–(4.36), the inverse inequalities $\|\xi_h\|_{0,F} \le C h_T^{-1/2}\|\xi_h\|_{0,T}$ and $\|\nabla\xi_h\|_{0,T} \le C h_T^{-1}\|\xi_h\|_{0,T}$, and Lemma 4.3,

$$\begin{aligned} & (\nabla\cdot(\mathbf{q}_h - \beta s_h),\xi_h)_{0,T} \\ =\ & -(\mathbf{q}_h - \beta s_h,\nabla\xi_h)_{0,T} + \sum_{F\in\mathcal{F}_T} ((\mathbf{q}_h - \beta s_h)\cdot n_T,\xi_h)_{0,F} \\ =\ & -(u_h - s_h,\beta\cdot\nabla\xi_h)_{0,T} + \sum_{F\in\mathcal{F}_T} (\beta\cdot n_T\{u_h\} + n_F^t n_F \gamma_{\beta,F}[\![u_h]\!] - \beta\cdot n_T s_h,\xi_h)_{0,F} \\ \le\ & C\|\beta\|_{\infty,T} h_T^{-1}\|\xi_h\|_{0,T} \sum_{F\in\widetilde{\mathcal{F}}_T} h_F^{1/2}\|[\![u_h]\!]\|_{0,F} \\ & + C h_T^{-1/2}\|\xi_h\|_{0,T} \sum_{F\in\mathcal{F}_T} \|\beta\cdot n_T\{u_h\} + n_F^t n_F \gamma_{\beta,F}[\![u_h]\!] - \beta\cdot n_T s_h\|_{0,F} \\ \le\ & C\|\beta\|_{\infty,T} h_T^{-1}\|\xi_h\|_{0,T} \sum_{F\in\widetilde{\mathcal{F}}_T} h_F^{1/2}\|[\![u_h]\!]\|_{0,F}, \end{aligned}$$

using finally the result proved previously for $\eta_{{\rm U},T}$, whence (4.47) follows.

According to Lemma 4.10, which holds true also in the general case, the estimate on $\eta_{{\rm DF},T}$ can be decomposed into three parts, the first two of which are bounded by

$$\tilde{C}\left(\frac{\lambda_{M,T}}{\lambda_{m,T}}\right)^{\frac{1}{2}} \|u - u_h\|_{B,*,\mathcal{F}_T}$$

with $\tilde{C}$ depending on d, k, $\kappa_{\mathcal{T}}$, α_F, θ, C_2, and C_3. The estimate for the third term is similar to that of [49] and can be obtained using the edge bubble function technique introduced in [98], yielding altogether (4.48).

Finally, the estimate (4.49) was established in [49] using the equivalence of norms on finite-dimensional spaces, inverse inequalities, and the definition of $\|\cdot\|_{B,T}$ by (4.5), following the approach given in [98]. □

Remark (Comments on the results of Theorem 4.18). In comparison with Theorem 4.17, again the crucial advantage of Theorem 4.18 is the confirmation of the localization of the error. However, the efficiency constant is no longer parameter-independent, the major overestimation being produced in the advection-dominated case. Nevertheless, as $h \to 0$, the estimators $\eta_{C,1,T}$, $\eta_{C,2,T}$, and $\eta_{U,T}$ will completely loose influence and $\eta_{NC,T}$ and $\eta_{DF,T}$ will become optimally efficient, the rapidity being a function of the local Péclet number $\frac{\|\beta\|_{\infty,T}}{\lambda_{m,T}} h_T$ on each $T \in \mathcal{T}_h$. This result is of the same quality as those achieved in [49, 98, 101, 102].

Remark (Efficiency of $\eta_{DF,T}$). In comparison with the results of [49, 98, 101, 102], the diffusive flux estimator $\eta_{DF,T}$ is not efficient with a constant of the form $c_1 + c_2 \min\{\mathrm{Pe}, \varrho\}$ with ϱ depending on β and K, but only of the form $c_1 + c_2\mathrm{Pe}$. The former efficiency can be obtained if integration by parts is performed and $\eta_{DF,T}$ is replaced by a minimum of $\eta_{DF,T}$ and an estimator as that in [49]. We did not perform here such a modification, also in view of the fact that the numerical experiments presented below show that $\eta_{DF,T}$ is actually small in comparison with the other estimators.

Remark (Efficiency of $\eta_{U,T}$ in comparison with finite volumes or mixed finite elements). In finite volume or mixed finite schemes, upwinding can likewise be used in order to stabilize the schemes in the advection-dominated regime. However, no efficiency of the corresponding upwinding estimator $\eta_{U,T}$ can be proved for these schemes, see [101] and [102], respectively. Contrarily to this situation, $\eta_{U,T}$ in the discontinuous Galerkin method is by (4.46) locally efficient (with a constant depending on the local Péclet number).

Remark (Efficiency of $\eta_{C,1,T}$ and $\eta_{U,T}$). We remark that the improvements in $\eta_{C,1,T}$ and $\eta_{U,T}$ described in Remark 4.6.4 were not taken into account in the proof of Theorem 4.18. Hence the actual efficiency of these estimators may be still better.

4.8 Numerical experiments

We present in this section the results of several numerical experiments.

4.8.1 Pure diffusion

For the pure diffusion problem we have examined three different test cases, all posed on the domain $\Omega = \{-1 < x, y < 1\}$ with Dirichlet boundary conditions. The diffusion tensor is isotropic (but heterogeneous in test cases 2 and 3) and can thus be identified with a scalar diffusion coefficient denoted by κ. The discrete solution has been obtained using

the weighted interior-penalty DG scheme proposed in [51] with polynomial degree $p = 1$, given by (4.9)–(4.11) with the penalty parameter and the weights given by (4.25)–(4.27). The diffusive flux $\mathbf{t}_h$ was reconstructed using (4.14)–(4.15) for both $l = 0$ or $l = 1$. Next, the piecewise affine Oswald interpolate $\mathcal{I}_{\rm Os}(u_h)$ of the discrete solution u_h was used. In the subsequent tables, sequences of uniformly refined, structured or unstructured meshes are considered to evaluate the convergence rates and N indicates the number of mesh elements. Columns labeled "eff" report the global effectivity index, that is the ratio of the a posteriori error estimate to the actual error, both quantities being computed over all mesh elements. We employ the following notation for the various error estimators: $\eta_{\rm NC} := \{\sum_{T\in\mathcal{T}_h} \eta_{{\rm NC},T}^2\}^{1/2}$, $\eta_{\rm R} := \{\sum_{T\in\mathcal{T}_h} \eta_{{\rm R},T}^2\}^{1/2}$, $\eta_{\rm DF} := \{\sum_{T\in\mathcal{T}_h} \eta_{{\rm DF},T}^2\}^{1/2}$, and so on.

For test case 1 the exact solution is $u(x,y) = \cos(0.5\pi x)\cos(0.5\pi y)$ and κ is equal to unity. The purpose of this test case is to assess the convergence rate of all the estimators in the case of a smooth solution. Tables 4.1 and 4.2 report the results obtained on structured and unstructured meshes, respectively. As expected, the residual estimator $\eta_{\rm R}$ converges to order $(l+2)$, *i.e.*, super-converges with respect to the nonconformity estimator and to the diffusive flux estimator, the latter always dominating the former by a factor between 2 and 3. For $l = 0$, the efficiency index is equal to 1.2. This exceptionally good result is actually below the value derived for the global efficiency of the abstract estimate in Theorem 4.6, namely $\sqrt{2}$. This is not a contradiction since in the present case, it turns out that the Oswald interpolate $\mathcal{I}_{\rm Os}(u_h)$ is closer to the discrete solution u_h than the exact solution u. For $l = 1$, the effectivity index is equal to 1.5 on structured meshes and to 1.3 on unstructured meshes, which confirms the sharpness of the estimate for $l = 1$ also. The effectivity index for $l = 1$ is however slightly larger than for $l = 0$, showing that for the present test case, the lowest-order diffusive flux reconstruction yields the sharpest results (a different conclusion is reached in the two following test cases).

			$l=0$			$l=1$		
N	$\|u-u_h\|_B$	$\eta_{\rm NC}$	$\eta_{\rm R}$	$\eta_{\rm DF}$	eff.	$\eta_{\rm R}$	$\eta_{\rm DF}$	eff.
128	3.28e-1	1.89e-1	7.23e-2	3.38e-1	1.2	5.50e-3	4.32e-1	1.4
512	1.62e-1	9.72e-2	1.82e-2	1.69e-1	1.2	6.90e-4	2.22e-1	1.5
2048	8.04e-2	4.89e-2	4.54e-3	8.39e-2	1.2	8.64e-5	1.12e-1	1.5
8192	4.01e-2	2.45e-2	1.14e-3	4.18e-2	1.2	1.08e-5	5.64e-2	1.5
order	1.0	1.0	2.0	1.0	-	3.0	1.0	-

Table 4.1: Convergence rates of error estimators for test case 1, structured meshes

			$l=0$			$l=1$		
N	$\|u-u_h\|_B$	$\eta_{\rm NC}$	$\eta_{\rm R}$	$\eta_{\rm DF}$	eff.	$\eta_{\rm R}$	$\eta_{\rm DF}$	eff.
112	3.16e-1	1.25e-1	7.01e-2	3.60e-1	1.2	5.13e-3	3.58e-1	1.2
448	1.58e-1	6.85e-2	1.76e-2	1.82e-1	1.2	6.90e-4	2.22e-1	1.5
1792	7.88e-2	3.53e-2	4.40e-3	9.10e-2	1.2	8.05e-5	9.43e-2	1.3
7168	3.93e-2	1.77e-2	1.10e-3	4.55e-2	1.2	1.01e-5	4.76e-2	1.3
order	1.1	1.1	2.1	1.1	-	3.2	1.1	-

Table 4.2: Convergence rates of error estimators for test case 1, unstructured meshes

Before moving to the following test cases, it is useful to compare the present error estimators to those previously available in the literature. We focus here on the classical error estimator for the pure diffusion case which consists of four terms: the nonconformity estimator (evaluated using the Oswald interpolate as in this paper) and three additional terms, namely the residual estimator $\eta_{\rm R}^*$, the diffusive flux (mass balance) estimator $\eta_{\rm DF}^*$, and the jump estimator $\eta_{\rm J}^*$ defined as follows:

$$(\eta_{\rm R}^*)^2 = \sum_{T\in\mathcal{T}_h} m_{T,K}^2 \|f+\nabla\cdot(K\nabla u_h)\|_{0,T}^2,$$
$$(\eta_{\rm DF}^*)^2 = \sum_{T\in\mathcal{T}_h} C_T \frac{h_T}{\lambda_{m,T}} \|n_F^t[\![K\nabla_h u_h]\!]\|_{0,\partial T\setminus\partial\Omega}^2,$$
$$(\eta_{\rm J}^*)^2 = \sum_{T\in\mathcal{T}_h} C_T \frac{1}{h_T} \|{\gamma_{K,F}}^{1/2}[\![u_h]\!]\|_{0,\partial T}^2,$$

where $C_T = 3dh_T|\partial T|/|T|$. Results are presented in Table 4.3. In particular, the 5th column, which displays the effectivity index, shows that the error is overestimated by a factor of 10. It should be observed that the main source for overestimation is the residual estimator $\eta_{\rm R}^*$, which we have transformed into a super-convergent term in the present work. The diffusive flux estimator $\eta_{\rm DF}^*$ is also observed to be about three-times larger than the estimators $\eta_{\rm DF}$ evaluated using the present reconstructed flux $\mathbf{t}_h$ either with $l=0$ or with $l=1$. For completeness, the last column of Table 4.3 proposes a comparison with the recent results of [49][6] where the mean value is subtracted from the residue within each mesh element, while the diffusive flux estimator and the nonconformity estimator are considered together as one unique term. We see that the estimate becomes sharper, though it still

[6] c'est-à-dire les résultats obtenus au chapitre 3 de cette thèse

overestimates the error by a factor of 4.5, mainly because the reconstructed flux $\mathbf{t}_h$ is not used.

N	η^*_{R}	η^*_{DF}	η^*_{J}	eff.	eff. [49]
112	1.74	1.38	3.57e-1	7.5	5.2
448	8.73e-1	5.86e-1	2.03e-1	7.2	4.9
1792	4.37e-1	2.75e-1	1.07e-1	7.1	4.7
7168	2.19e-1	1.31e-1	5.42e-2	7.1	4.5
order	1.1	1.1	1.0	-	-

Table 4.3: Comparison with other error estimators for test case 1, unstructured meshes

The aim of test cases 2 and 3, which were proposed in [88], is to address the question of diffusion heterogeneities. The domain Ω is split along the Cartesian axes into four subregions Ω_i. The subregion $\{x > 0, y > 0\} \cap \Omega$ is indicated by Ω_1 and the subsequent numbering is done in a counterclockwise manner. The diffusion coefficient is equal to κ_i in subregion Ω_i where κ_i is a constant. Taking the forcing term equal to zero, the analytical solution with corresponding nonhomogeneous Dirichlet boundary conditions can be written in polar coordinates as

$$u(r,\phi)|_{\Omega_i} = r^\alpha \left(a_i \sin(\alpha\phi) + b_i \cos(\alpha\phi)\right),$$

where the subscript i refers to the corresponding subregion. Owing to the singularity in the origin, the calculated solution converges with order 2α in the L^2-norm and with order α in the energy (semi-)norm. For test case 2, we take $\kappa_1 = \kappa_3 = 5$ and $\kappa_2 = \kappa_4 = 1$, yielding $\alpha = 0.53544095$ and

$$\begin{aligned} a_1 &= \ \ 0.44721360; & b_1 &= \ \ 1.00000000; \\ a_2 &= -0.74535599; & b_2 &= \ \ 2.33333333; \\ a_3 &= -0.94411759; & b_3 &= \ \ 0.55555556; \\ a_4 &= -2.40170264; & b_4 &= -0.48148148. \end{aligned}$$

For test case 3, we take $\kappa_1 = \kappa_3 = 100$ and $\kappa_2 = \kappa_4 = 1$, yielding $\alpha = 0.12690207$ and

$$\begin{aligned} a_1 &= \ 0.10000000; & b_1 &= \ 1.00000000; \\ a_2 &= -9.60396040; & b_2 &= \ 2.96039604; \\ a_3 &= -0.48035487; & b_3 &= -0.88275659; \\ a_4 &= \ 7.70156488; & b_4 &= -6.45646175. \end{aligned}$$

The results for test case 2 are shown in Tables 4.4 and 4.5 for structured and unstructured meshes, respectively. Since the forcing term is zero, the residual estimator is also equal to zero, and has not been reported. The interpolation error on nonhomogeneous Dirichlet boundary conditions is not reported either. We observe that the expected convergence rate of order α is obtained for both the nonconformity estimator η_{NC} and for the diffusive flux estimator η_{DF}. Both estimators yield comparable values. The effectivity index is 1.9 for $l = 0$ and 1.8 for $l = 1$; hence, for this test case, employing $l = 1$ for the reconstruction leads to a slightly sharper estimator.

			$l=0$		$l=1$	
N	$\|u-u_h\|_B$	η_{NC}	η_{DF}	eff.	η_{DF}	eff.
128	6.61e-01	9.60e-1	8.02e-1	1.9	6.54e-1	1.8
512	4.58e-01	6.68e-1	5.63e-1	1.9	4.63e-1	1.8
2048	3.17e-01	4.62e-1	3.92e-1	1.9	3.23e-1	1.8
8192	2.19e-01	3.20e-1	2.72e-1	1.9	2.25e-1	1.8
order	0.53	0.53	0.53	-	0.53	-

Table 4.4: Convergence rates of error estimators for test case 2, structured meshes

			$l=0$		$l=1$	
N	$\|u-u_h\|_B$	η_{NC}	η_{DF}	eff.	η_{DF}	eff.
112	6.11e-01	8.70e-1	7.43e-1	1.9	6.00e-1	1.7
448	4.28e-01	6.09e-1	5.35e-1	1.9	4.32e-1	1.7
1792	2.97e-01	4.23e-1	3.74e-1	1.9	3.05e-1	1.8
7168	2.01e-01	2.92e-1	2.60e-1	1.9	2.12e-1	1.8
order	0.53	0.53	0.53	-	0.52	-

Table 4.5: Convergence rates of error estimators for test case 2, unstructured meshes

The results for test case 3 are shown in Tables 4.6 and 4.7 for structured and unstructured meshes, respectively. The order of convergence of the error estimators is close to α, and the error is overestimated by a factor of approximately 3.8. This is because the nonconformity error estimator now dominates over the diffusive flux estimator. Hence, although the diffusive flux estimator is lower for $l = 1$ than for $l = 0$, this difference is scarcely reflected in the effectivity index. Finally, it is worthwhile to notice that in the present setting, the diffusion coefficient is not monotone around the singularity, thus precluding the use of weighted variants of the Oswald interpolate such as that proposed in [3]. On the other hand, one can employ a piecewise quadratic Oswald interpolate as in [101,102].

			$l=0$		$l=1$	
N	$\|u-u_h\|_B$	η_{NC}	η_{DF}	eff.	η_{DF}	eff.
128	3.49	12.4	2.68	3.6	2.02	3.6
512	3.29	11.9	2.57	3.7	1.95	3.6
2048	3.09	11.3	2.45	3.7	1.86	3.7
8192	2.88	10.7	2.32	3.8	1.76	3.8
order	0.10	0.08	0.08	-	0.08	-

Table 4.6: Convergence rates of error estimators for test case 3, structured meshes

			$l=0$		$l=1$	
N	$\|u-u_h\|_B$	η_{NC}	η_{DF}	eff.	η_{DF}	eff.
112	3.27	11.8	2.39	3.7	1.89	3.7
448	3.11	11.3	2.33	3.7	1.84	3.7
1792	2.93	10.8	2.23	3.8	1.77	3.7
7168	2.75	10.3	2.12	3.8	1.68	3.8
order	0.09	0.08	0.08	-	0.07	-

Table 4.7: Convergence rates of error estimators for test case 3, unstructured meshes

4.8.2 Advection–diffusion–reaction

We consider the domain $\Omega = \{0 < x, y < 1\}$, the reaction coefficient $\mu = 1$, the advection field $\beta = (1, 0)^t$, and an isotropic homogeneous diffusion tensor represented by a diffusion coefficient κ. We run tests with $\kappa = 10^{-2}$ (test case 4) and $\kappa = 10^{-4}$ (test case 5). The

source term f is chosen so that the exact solution with homogeneous Dirichlet boundary conditions is

$$u = \tfrac{1}{2}x(x-1)y(y-1)\left(1-\tanh(10-20x)\right).$$

For brevity, only results for uniformly refined structured meshes are presented.

N	$\|u-u_h\|_B$	η_{NC}	eff. $l=0$	eff. $l=1$
128	1.95e-3	3.62e-3	13.8	14.4
512	4.01e-4	1.84e-3	11.1	10.9
2048	1.89e-3	8.84e-4	8.10	7.75
order	1.1	1.1	-	-

Table 4.8: Efficiency of error estimators for test case 4 ($\kappa = 10^{-2}$)

		$l=0$			$l=1$			
N	η_{R}^*	η_{R}	η_{DF}	η_{U}	η_{R}	η_{DF}	η_{U}	$\eta_{\mathrm{C},1}$
128	4.91e-2	3.94e-2	8.82e-3	6.35e-2	1.12e-2	8.73e-3	6.35e-2	3.28e-2
512	1.44e-2	9.86e-3	4.93e-3	2.87e-2	1.66e-3	4.73e-3	2.87e-2	7.69e-3
2048	4.63e-3	2.42e-3	2.51e-3	9.77e-3	3.19e-4	2.37e-3	9.77e-3	1.53e-3
order	1.6	2.0	1.0	1.6	2.4	1.0	1.6	2.3

Table 4.9: Convergence of error estimators for test case 4 ($\kappa = 10^{-2}$)

Tables 4.8 and 4.9 report the results for $\kappa = 10^{-2}$. Table 4.8 focuses on the global effectivity index when both the diffusive and convective fluxes are reconstructed using $l = 0$ or $l = 1$. Both choices yield comparable results with efficiency indices ranging between 8 and 14 approximately. A more detailed comparison can be found in Table 4.9. The residual estimator η_{R} super-converges and converges faster for $l = 1$ than for $l = 0$. The classical residual estimator η_{R}^* evaluated using solely the discrete solution is also reported; it takes, as expected, larger values. The diffusive flux estimator η_{DF} yields the smallest contribution among the different terms in the error estimate. The upwinding estimator η_{U} is dominant, along with the first advection estimator $\eta_{\mathrm{C},1}$ for $l = 1$, while this latter estimator vanishes for $l = 0$ since in this case, $\nabla\cdot(\mathbf{q}_h - \beta\mathcal{I}_{\mathrm{Os}}(u_h))$ is by construction piecewise constant. Finally, the second advection estimator $\eta_{\mathrm{C},2}$ vanishes identically because β is divergence-free. All in all, there is little gain in efficiency when going from $l = 0$ to $l = 1$.

N	$\|u-u_h\|_B$	η_{NC}	eff. $l=0$	eff. $l=1$
128	1.72e-3	2.73e-3	80	89
512	5.68e-4	6.74e-4	124	128
order	1.4	2.0	-	-

Table 4.10: Efficiency of error estimators for test case 5 ($\kappa = 10^{-4}$)

		$l=0$			$l=1$			
N	η_{R}^*	η_{R}	η_{DF}	η_{U}	η_{R}	η_{DF}	η_{U}	$\eta_{\mathrm{C},1}$
128	7.77e-2	6.84e-2	1.06e-3	6.98e-2	1.92e-2	1.03e-3	6.98e-2	6.55e-2
512	3.90e-2	3.41e-2	6.20e-4	3.60e-2	3.44e-3	5.71e-4	3.60e-2	3.38e-2
order	1.1	1.0	0.8	1.0	2.5	0.8	1.0	1.0

Table 4.11: Convergence of error estimators for test case 5 ($\kappa = 10^{-4}$)

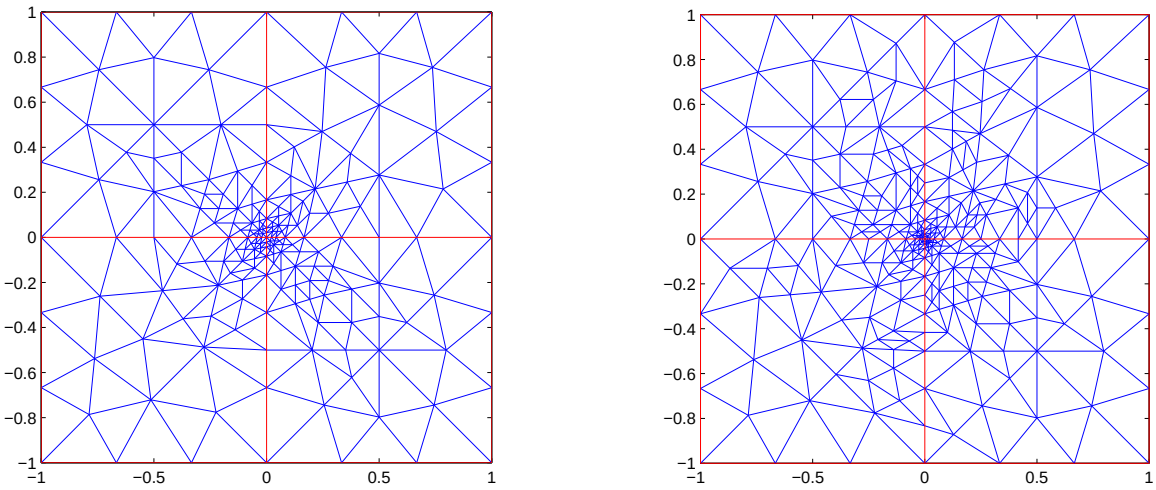
Tables 4.10 and 4.11 report the results for $\kappa = 10^{-4}$. Table 4.10 focuses on the global effectivity index for the $l = 0$ and $l = 1$ flux reconstructions. Again, both choices yield similar results, and the efficiency indices are roughly ten-times larger than those observed for $\kappa = 10^{-2}$, in agreement with the cut-off coefficients employed in front of the estimators. A more detailed comparison can be found in Table 4.11. As for test case 4, the residual estimator η_{R} converges faster for $l = 1$ than for $l = 0$, but this gain is compensated by the first advection estimator $\eta_{\mathrm{C},1}$. The diffusive flux estimator η_{DF} yields the smallest contribution, while the upwinding estimator η_{U} dominates the overall estimate.

4.8.3 Adaptive meshes

We conclude this section by an example on how the error estimator with $l = 0$ can be used to adapt the mesh. Test case 2 is considered. The adaptive mesh refinement procedure flags 5% of the mesh elements yielding the largest error indicators. Results are reported in Table 4.12. The effectivity index fluctuates between 1.7 and 2 and decreases as finer meshes are constructed. Comparing with Table 4.5, we observe that the energy norm of the error on an adapted mesh with 494 elements is comparable to that obtained on a uniformly refined mesh with 7168 elements. Finally, Figure 4.1 presents two meshes obtained within the adaptive refinement procedure, one with 342 elements and one with 494 elements. We see that the adaptive refinement correctly aims at capturing the singularity at the origin.

N	$\|u - u_h\|_B$	η_{NC}	η_{DF}	eff.
112	6.11e-1	8.70e-1	7.43e-1	1.87
148	4.58e-1	6.17e-1	5.78e-1	1.84
204	3.59e-1	5.59e-1	4.63e-1	2.02
264	2.96e-1	4.21e-1	3.76e-1	1.91
342	2.50e-1	3.05e-1	3.23e-1	1.78
494	2.10e-1	2.20e-1	2.78e-1	1.68

Table 4.12: Error as a function of mesh elements



342 elements

494 elements

Figure 4.1: Two meshes successively refined using the error estimator with $l = 0$ reconstruction: 342 elements (left) and 494 elements (right)

Chapitre 5

Conclusions et perspectives

Dans ce mémoire, nous nous sommes intéressés aux méthodes de Galerkine discontinues et à l'analyse *a posteriori* pour les équations d'advection-diffusion-réaction linéaires et stationnaires avec diffusion hétérogène.

Dans le chapitre 2, nous avons présenté une méthode de Galerkine discontinue que nous avons nommée SWIP. La méthode est une variante de la méthode SIPG, et utilise des poids pondérés qui sont calculés en fonction du tenseur de diffusion. Le terme de pénalisation dépend de la moyenne arithmétique de la diffusion dans la direction normale à l'interface. Avec l'analyse d'erreur *a priori* de la méthode nous avons montré que la convergence est optimale en le pas du maillage et qu'elle est indépendante des hétérogénéités de la diffusion dans la norme d'énergie et la norme L^2 (sous hypothèse de régularité elliptique). Par contre, la convergence dans la norme advective peut être influencée par l'anisotropie locale du tenseur de diffusion. Les tests numériques que nous avons effectués confirment les résultats théoriques. Une comparaison avec la méthode SIPG montre que la méthode SWIP offre une alternative intéressante quand une diffusion localement petite est à l'origine d'une couche limite qui n'est pas suffisamment résolue par le maillage.

Dans le chapitre 3, nous avons présenté un premier estimateur d'erreur *a posteriori* pour la semi-norme d'énergie. L'estimateur d'erreur est intégralement calculable en utilisant la solution calculée, les données du problème et le maillage, et a été obtenu en effectuant un analyse d'erreur *a posteriori* par résidus. Dans l'analyse nous avons distingué entre l'erreur directement associée au résidu sur chaque maille, l'erreur de non-conformité des flux diffusifs et l'erreur de non-conformité de la solution elle-même. Nous avons montré que les deux premiers termes ne dépendent pas des hétérogénéités du tenseur de diffusion ; ce résultat a été obtenu grâce aux poids pondérés utilisés dans la méthode SWIP. Les tests

numériques sont en accord avec la théorie, même lorsque la solution exacte est singulière du fait des hétérogénéités de la diffusion.

Dans le chapitre 4, nous avons présenté un deuxième estimateur d'erreur *a posteriori*, cette fois obtenu avec des champs vectoriels auxiliaires qui appartiennent à l'espace des éléments finis de Raviart-Thomas-Nédélec. Les champs vectoriels sont obtenus en résolvant des problèmes locaux. Nous avons montré que, sous certaines hypothèses sur le coefficient de réaction et la divergence du champ advectif, le terme de résidu est de la forme $\|f - \pi_k f\|$, où π_k indique la projection L^2 orthogonale sur l'espace vectoriel des polynômes de degré inférieur ou égal à k. Les tests numériques ont confirmé la bonne convergence des estimateurs, et l'indice d'efficacité est en accord avec la théorie. En particulier, l'indice d'efficacité est meilleur que celui obtenu au chapitre 3. L'estimateur obtenu au chapitre 4 est particulièrement utile pour l'adaptation de maillage.

La théorie et les tests numériques ont montré que l'erreur de non-conformité calculée avec l'interpolé de Oswald est influencée par les hétérogénéités du tenseur de diffusion. Pour améliorer l'indice d'efficacité des deux estimateurs d'erreur présentés dans ce mémoire, il serait intéressant de pouvoir évaluer l'erreur de non-conformité de manière indépendante des hétérogénéités de la diffusion. Par ailleurs, une prochaine étape importante serait d'étendre ce travail à l'équation d'advection-diffusion-réaction instationnaire. Il faudrait d'abord choisir une discrétisation en temps pour le schéma SWIP. Une extension des estimateurs d'erreur dans ce cadre devrait idéalement distinguer entre les erreurs dues à la discrétisation en temps et celles dues à la discrétisation en espace, afin d'indiquer s'il convient d'agir sur le maillage ou sur le pas de temps. Afin d'obtenir de bonnes solutions approchées sans un coût de calcul excessif, il faudrait enfin élaborer et tester différentes stratégies d'adaptation de maillage.

Bibliographie

[1] B. Achchab, S. Achchab, A. Agouzal, and R. Ellaia. On a posteriori error estimator for primal, equilibrium and mixed approximation of diffusion equations. *Appl. Math. Comput.*, 134(1) :83–92, 2003.

[2] Y. Achdou, C. Bernardi, and F. Coquel. A priori and a posteriori analysis of finite volume discretizations of Darcy's equations. *Numer. Math.*, 96 :17–42, 2003.

[3] M. Ainsworth. Robust a posteriori error estimation for nonconforming finite element approximation. *SIAM J. Numer. Anal.*, 42(6) :2320–2341, 2005.

[4] M. Ainsworth. A synthesis of a posteriori error estimation techniques for conforming, non-conforming and discontinuous Galerkin finite element methods. In *Recent advances in adaptive computation*, volume 383 of *Contemp. Math.*, pages 1–14. Amer. Math. Soc., Providence, RI, 2005.

[5] M. Ainsworth. A posteriori error estimation for discontinuous Galerkin finite element approximation. *SIAM J. Numer. Anal.*, 45(4) :1777–1798, 2007.

[6] V. Aizinger and C. Dawson. The local discontinuous Galerkin method for three-dimensional shallow water flow. *Comput. Methods Appl. Mech. Engrg.*, 196(4-6) :734–746, 2007.

[7] M. Anderson. *Groundwater Contamination.* National Academy Press, Washington, DC 20418, 1984.

[8] D. N. Arnold, F. Brezzi, B. Cockburn, and D. Marini. Discontinuous Galerkin methods for elliptic problems. In *Discontinuous Galerkin Methods : Theory, Computation and Applications*, pages 89–101, 2000.

[9] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal*, 39(5) :1749–1779, 2001/02.

[10] D.N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 19 :742–760, 1982.

[11] I. Babuška. The finite element method with penalty. *Math. Comp.*, 27 :221–228, 1973.

[12] I. Babuška and W. Rheinbolt. Error estimates for adaptive finite element method computations. *SIAM J. Numer. Anal.*, 15 :736–754, 1978.

[13] I. Babuška and W. Rheinbolt. A posteriori error estimates for the finite element method. *Int. J. Numer. Methods Engrg.*, 12 :1597–1615, 1978.

[14] I. Babuška and M. Zlámal. Nonconforming elements in the finite element method with penalty. *SIAM J. Numer. Anal.*, 10 :863–875, 1973.

[15] G. A. Baker. Finite element methods for elliptic equations using nonconforming elements. *Math. Comp.*, 31 :45–59, 1977.

[16] F. Bassi and S. Rebay. A high–order accurate dicontinuous finite element method for numerical solution of the compressible Nvier–Stokes equations. *J. Comput. Phys.*, 131 :267–279, 1997.

[17] F. Bassi, S. Rebay, G. Mariotti, S. Pedinotti, and M. Savini. A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows. In R. Decypere and G. Dibelius, editors, *Proceedings of 2nd European Conference on Turbomachinery, Fluid Dynamics and Thermodynamics*, pages 99–108, Antwerpen, Belgium, 1997. Technologish Instituut.

[18] P. Bastian and B. Rivière. Superconvergence and $H(\mathrm{div})$ projection for discontinuous Galerkin methods. *Internat. J. Numer. Methods Fluids*, 42(10) :1043–1057, 2003.

[19] M. Bebendorf. A note on the Poincaré inequality for convex domains. *Z. Anal. Anwendungen*, 22(4) :751–756, 2003.

[20] R. Becker, P. Hansbo, and M. G. Larson. Energy norm a posteriori error estimation for discontinuous Galerkin methods. *Comput. Methods Appl. Mech. Engrg.*, 192(5-6) :723–733, 2003.

[21] R. Becker, P. Hansbo, and R. Stenberg. A finite element method for domain decomposition with non-matching grids. *M2AN Math. Model. Numer. Anal.*, 37(2) :209–225, 2003.

[22] C. Bernardi and R. Verfürth. Adaptive finite element methods for elliptic equations with non-smooth coefficients. *Numer. Math.*, 85(4) :579–608, 2000.

[23] F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods*, volume 15 of *Springer Series in Computational Mathematics*. Springer-Verlag, New York, 1991.

[24] E. Burman and A. Ern. Continuous interior penalty *hp*-finite element methods for advection and advection-diffusion equations. *Math. Comp.*, 76(259) :1119–1140, 2007.

[25] E. Burman and P. Zunino. A domain decomposition method based on weighted interior penalties for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.*, 44(4) :1612–1638, 2006.

[26] R. Bustinza, G. N. Gatica, and B. Cockburn. An a posteriori error estimate for the local discontinuous Galerkin method applied to linear and nonlinear diffusion problems. *J. Sci. Comput.*, 22/23 :147–185, 2005.

[27] C. Carstensen, S. Bartels, and S. Jansche. A posteriori error estimates for nonconforming finite element methods. *Numer. Math.*, 92(2) :233–256, 2002.

[28] C. Carstensen and S. A. Funken. Constants in Clément-interpolation error and residual based a posteriori error estimates in finite element methods. *East-West J. Numer. Math.*, 8(3) :153–175, 2000.

[29] P. Castillo. An a posteriori error estimate for the local discontinuous Galerkin method. *J. Sci. Comput.*, 22/23 :187–204, 2005.

[30] S. Cochez-Dhondt and S Nicaise. Equilibrated error estimators for discontinuous Galerkin methods. *Submitted*, 2007.

[31] B. Cockburn, G. E. Karniadakis, and C.-W. Shu, editors. *Discontinuous Galerkin methods*, volume 11 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, Berlin, 2000. Theory, computation and applications, Papers from the 1st International Symposium held in Newport, RI, May 24–26, 1999.

[32] B. Cockburn and C.-W. Shu. The Runge-Kutta local projection P^1-discontinuous-Galerkin finite element method for scalar conservation laws. *RAIRO Modél. Math. Anal. Numér.*, 25(3) :337–361, 1991.

[33] B. Cockburn and C.-W. Shu. The local discontinuous Galerkin method for time-dependent convection-diffusion systems. *SIAM, J. Numer. Anal.*, 35 :2440–2463, 1998.

[34] J.-P. Croisille, A. Ern, T. Lelièvre, and J. Proft. Analysis and simulation of a coupled hyperbolic/parabolic model problem. *J. Numer. Math.*, 13(2) :81–103, 2005.

[35] E. Dari, R. Duran, C. Padra, and V. Vampa. A posteriori error estimators for nonconforming finite element methods. *RAIRO Modél. Math. Anal. Numér.*, 30(4) :385–400, 1996.

[36] L. M. Delves and C. A. Hall. An implicit matching principle for global element calculations. *J. Inst. Math. Appl.*, 23(2) :223–234, 1979.

[37] P. Destuynder and B. Métivet. Explicit error bounds in a conforming finite element method. *Math. Comp.*, 68(228) :1379–1396, 1999.

[38] D. A. Di Pietro, A. Ern, and J.-L. Guermond. Discontinuous Galerkin methods for anisotropic diffusion with advection. *SIAM J. Numer. Anal.*, 2007. To appear.

[39] W. Dörfler. A convergent adaptive algorithm for Poisson's equation. *SIAM J. Numer. Anal.*, 33(3) :1106–1124, 1996.

[40] J. Douglas Jr. and T. Dupont. Interior penalty procedures for elliptic and parabolic Galerkin methods. In *Lecture Notes in Phys. 58*. Springer-Verlag, 1976.

[41] L. El Alaoui and A. Ern. Residual and hierarchical a posteriori error estimates for nonconforming mixed finite element methods. *M2AN Math. Model. Numer. Anal.*, 38(6) :903–929, 2004.

[42] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, NY, 2004.

[43] A. Ern and J.-L. Guermond. Discontinuous Galerkin methods for Friedrichs' systems. I. General theory. *SIAM J. Numer. Anal.*, 44(2) :753–778, 2006.

[44] A. Ern and J.-L. Guermond. Discontinuous Galerkin methods for Friedrichs' systems. II. Second-order elliptic PDEs. *SIAM J. Numer. Anal.*, 44(6) :2363–2388, 2006.

[45] A. Ern and J.-L. Guermond. Discontinuous Galerkin methods for Friedrichs' systems. Part III. Multi-field theories with partial coercivity. *SIAM J. Numer. Anal.*, 2007. To appear.

[46] A. Ern, S. Nicaise, and M. Vohralík. An accurate $\mathbf{H}(\mathrm{div})$ flux reconstruction for discontinuous Galerkin approximations of elliptic problems. *C. R. Math. Acad. Sci. Paris*, 2007. To appear.

[47] A. Ern, S. Piperno, and K. Djadel. A well-balanced runge–kutta discontinuous galerkin method for the shallow-water equations with flooding and drying. *Internat. J. Numer. Methods Fluids*, 2007. To appear.

[48] A. Ern and J. Proft. Multi-algorithmic methods for coupled hyperbolic-parabolic problems. *Int. J. Numer. Anal. Model.*, 1(3) :94–114, 2006.

[49] A. Ern and A. F. Stephansen. A posteriori energy-norm error estimates for advection-diffusion equations approximated by weighted interior penalty methods. Technical Report 364, CERMICS/ENPC, 2007.

[50] A. Ern, A. F. Stephansen, and M. Vohralík. Improved energy norm a posteriori error estimation based on flux reconstruction for discontinuous Galerkin methods. Technical report, CERMICS/ENPC, 2007.

[51] A. Ern, A. F. Stephansen, and P. Zunino. A discontinuous Galerkin method with weighted averages for advection-diffusion equations with locally vanishing and anisotropic diffusivity. *IMA J. Numer. Anal.*, 2007. Accepted.

[52] F. Gastaldi and A. Quarteroni. On the coupling of hyperbolic and parabolic systems : analytical and numerical approach. *Appl. Numer. Math.*, 6(1-2) :3–31, 1989/90. Spectral multi-domain methods (Paris, 1988).

[53] E. H. Georgoulis and A. Lasis. A note on the design of *hp*-version interior penalty Galerkin finite element methods *IMA J. Numer. Anal.*, 26(2) :381–390, 2006.

[54] B. Heinrich and S. Nicaise. The Nitsche mortar finite-element method for transmission problems with singularities. *IMA J. Numer. Anal.*, 23(2) :331–358, 2003.

[55] B. Heinrich and K. Pietsch. Nitsche type mortaring for some elliptic problem with corner singularities. *Computing*, 68(3) :217–238, 2002.

[56] B. Heinrich and K. Pönitz. Nitsche type mortaring for singularly perturbed reaction-diffusion problems. *Computing*, 75(4) :257–279, 2005.

[57] P. Houston, I. Perugia, and D. Schötzau. Energy norm a posteriori error estimation for mixed Galerkin approximation of the Maxwell operator. *Comput. Methods Appl. Mech. Engrg.*, 194 :499–510, 2005.

[58] P. Houston, D. Schötzau, and Th. P. Wihler. Energy norm a posteriori error estimation of *hp*-adaptive discontinuous Galerkin methods for elliptic problems. *Math. Models Methods Appl. Sci.*, 17(1) :33–62, 2007.

[59] P. Houston, Ch. Schwab, and E. Süli. Discontinuous *hp*-finite element methods for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.*, 39(6) :2133–2163, 2002.

[60] G. Jiang and C.-W. Shu. On a cell inequality for discontinuous Galerkin methods. *Math. Comp.*, 62(206) :531–538, 1994.

[61] C. Johnson and J. Pitkäranta. An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comp.*, 46(173) :1–26, 1986.

[62] O. A. Karakashian and F. Pascal. A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems. *SIAM J. Numer. Anal.*, 41(6) :2374–2399, 2003.

[63] K. Y. Kim. A posteriori error analysis for locally conservative mixed methods. *Math. Comp.*, 76(257) :43–66, 2007.

[64] K. Y. Kim. A posteriori error estimators for locally conservative methods of nonlinear elliptic problems. *Appl. Numer. Math.*, 57 :1065–1080, 2007.

[65] P. Ladevèze. *Comparaison de modèles de milieux continus.* PhD thesis, Université Pierre et Marie Curie, 1975.

[66] P. Ladevèze and D. Leguillon. Error estimate procedure in the finite element method and applications. *SIAM J. Numer. Anal.*, 20(3) :485–509, 1983.

[67] R. Lazarov, S. Repin, and S. Tomar. Functional a posteriori error estimates for discontinuous Galerkin approximations of elliptic problems. *Report 2006-40, Ricam, Austria*, 2006.

[68] P. Lesaint. *Sur la résolution des systèmes hyperboliques du premier ordre par des méthodes d'éléments finis.* PhD thesis, University of Paris VI, 1975.

[69] P. Lesaint and P.-A. Raviart. On a finite element method for solving the neutron transport equation. In C. de Boors, editor, *Mathematical aspects of Finite Elements in Partial Differential Equations*, pages 89–123. Academic Press, 1974.

[70] P. Monk and E. Süli. The adaptive computation of far-field patterns by a posteriori error estimation of linear functionals. *SIAM J. Numer. Anal.*, 36(1) :251–274, 1999.

[71] P. Morin, R. H. Nochetto, and K. G. Siebert. Data oscillation and convergence of adaptive FEM. *SIAM J. Numer. Anal.*, 38(2) :466–488, 2000.

[72] P. Morin, R. H. Nochetto, and K. G. Siebert. Convergence of adaptive finite element methods. *SIAM Rev.*, 44(4) :631–658 (2003), 2002. Revised reprint of "Data oscillation and convergence of adaptive FEM" [SIAM J. Numer. Anal. **38** (2000), no. 2, 466–488 ; MR1770058 (2001g :65157)].

[73] P. Morin, R. H. Nochetto, and K. G. Siebert. Local problems on stars : a posteriori error estimators, convergence, and performance. *Math. Comp.*, 72(243) :1067–1097, 2003.

[74] J.-C. Nédélec. Mixed finite elements in $\mathbb{R}^3$. *Numer. Math.*, 35(3) :315–341, 1980.

[75] P. Neittaanmäki and S. Repin. *Reliable methods for computer simulation*, volume 33 of *Studies in Mathematics and its Applications*. Elsevier Science B.V., Amsterdam, 2004. Error control and a posteriori estimates.

[76] J. Nitsche. Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind. *Abh. Math. Sem. Univ. Hamburg*, 36 :9–15, 1971. Collection of articles dedicated to Lothar Collatz on his sixtieth birthday.

[77] J. Nitsche. On Dirichlet problems using subspaces with nearly zero boundary conditions. In *The mathematical foundations of the finite element method with applications to partial differential equations (Proc. Sympos., Univ. Maryland, Baltimore, Md., 1972)*, pages 603–627. Academic Press, New York, 1972.

[78] J. T. Oden, I. Babuška, and C. E. Baumann. A discontinuous *hp* finite element method for diffusion problems. *J. Comput. Phys.*, 146(2) :491–519, 1998.

[79] L. E. Payne and H. F. Weinberger. An optimal Poincaré inequality for convex domains. *Arch. Rational Mech. Anal.*, 5 :286–292 (1960), 1960.

[80] M. Petzoldt. Regularity results for Laplace interface problems in two dimensions. *Z. Anal. Anwendungen*, 20(2) :431–455, 2001.

[81] W. Prager and J. L. Synge. Approximations in elasticity based on the concept of function space. *Quart. Appl. Math.*, 5 :241–269, 1947.

[82] P.-A. Raviart and J.-M. Thomas. A mixed finite element method for 2nd order elliptic problems. In *Mathematical aspects of finite element methods (Proc. Conf., Consiglio Naz. delle Ricerche (C.N.R.), Rome, 1975)*, pages 292–315. Lecture Notes in Math., Vol. 606. Springer, Berlin, 1977.

[83] W.H. Reed and T.R. Hill. Triangular mesh methods for the neutron transport equation. Technical Report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.

[84] S. I. Repin. A posteriori error estimation for nonlinear variational problems by duality theory. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, 243(Kraev. Zadachi Mat. Fiz. i Smezh. Vopr. Teor. Funktsii. 28) :201–214, 342, 1997.

[85] S. I. Repin. A unified approach to a posteriori error estimation based on duality error majorants. *Math. Comput. Simulation*, 50(1-4) :305–321, 1999. Modelling '98 (Prague).

[86] S. I. Repin. A posteriori error estimation for variational problems with uniformly convex functionals. *Math. Comp.*, 69(230) :481–500, 2000.

[87] B. Rivière and M. F. Wheeler. A posteriori error estimates for a discontinuous Galerkin method applied to elliptic problems. *Comput. Math. Appl.*, 46(1) :141–163, 2003.

[88] B. Rivière, M. F. Wheeler, and K. Banas. Part II. Discontinuous Galerkin method applied to single phase flow in porous media. *Comput. Geosci.*, 4(4) :337–349, 2000.

[89] B. Rivière, M. F. Wheeler, and V. Girault. Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. Part I. *Comput. Geosci.*, 3 :337–360, 1999.

[90] B. Rivière, M. F. Wheeler, and V. Girault. A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems. *SIAM J. Numer. Anal.*, 39(3) :902–931, 2001.

[91] J. E. Roberts and J.-M. Thomas. Mixed and hybrid methods. In *Handbook of Numerical Analysis, Vol. II*, pages 523–639. North-Holland, Amsterdam, 1991.

[92] A. Romkes, J. T. Oden, and S. Prudhomme. A priori error analyses of a stabilized discontinuous Galerkin method. ICES Technical Report TICAM-02-28, University of Texas, july 2002.

[93] R. Schneider, Y. Xu, and A. Zhou. An analysis of discontinuous Galerkin methods for elliptic problems. *Adv. Comput. Math.*, 25(1-3) :259–286, 2006.

[94] R. Stenberg. Mortaring by a method of J.A. Nitsche. In Idelsohn S.R., Oñate E., and Dvorkin E.N., editors, *Computational Mechanics : New trends and applications*, Barcelona, Spain, 1998.

[95] P. A. Tassi, O. Bokhove, and C. A. Vionnet. Space discontinuous Galerkin method for shallow water flows -kinetic and hllc flux, and potential vorticity generation. *Adv. in Water Res.*, 30(4) :998–1015, 2007.

[96] R. Verfürth. A posteriori error estimations and adaptative mesh-refinement techniques. *J. Comput. Appl. Math.*, 50 :67–83, 1994.

[97] R. Verfürth. *A Review of a Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques.* Wiley, Chichester, UK, 1996.

[98] R. Verfürth. A posteriori error estimators for convection-diffusion equations. *Numer. Math.*, 80(4) :641–663, 1998.

[99] R. Verfürth. Robust a posteriori error estimates for stationary convection–diffusion equations. *Siam J. Numer. Anal.*, 43(5) :1783–1802, 2005.

[100] M. Vohralík. On the discrete Poincaré-Friedrichs inequalities for nonconforming approximations of the Sobolev space H^1. *Numer. Funct. Anal. Optim.*, 26(7-8) :925–952, 2005.

[101] M. Vohralík. Residual flux-based a posteriori error estimates for finite volume discretizations of inhomogeneous, anisotropic, and convection-dominated problems. *Submitted*, 2006.

[102] M. Vohralík. A posteriori error estimates for lowest-order mixed finite element discretizations of convection–diffusion–reaction equations. *SIAM J. Numer. Anal.*, 45(4) :1570–1599, 2007.

[103] M. Vohralík. Unified primal formulation-based a priori and a posteriori error analysis of mixed finite element methods. *In preparation*, 2007.

[104] M. F. Wheeler. An elliptic collocation-finite element method with interior penalties. *SIAM J. Numer. Anal.*, 15 :152–161, 1978.